SI 201: Discussion 6

Working with CSV files and Midterm Review

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

CSV Format

- **CSV (comma separated values) files are a simple and lightweight way to store structured data**. They can be read by many different programs and are a common format for sharing datasets.

- A CSV file represents data as a **series of rows and columns**, much like an Excel spreadsheet or a matrix.

CSV Format & Example CSV data

- First row of a CSV: Usually contains the name of each column. This is called the **header row**. Most CSVs have one, but it isn't required.
- Each line of a CSV: Represents a row. Usually, the columns are separated from each other using commas (,) within each row, unless specified otherwise.
- Other separators such as tabs (\t) and pipes (|) can be used, but commas (,) are by far the most common.

```
month,date,sample,Harris result,Trump result
sept,19,1880 LV,0.51,0.45
sept,19,1880 LV,0.53,0.47
sept,17,810 LV,0.49,0.45
sept,17,820 RV,0.49,0.45
sept,17,1445 RV,0.49,0.45
sept,17,1000 LV,0.53,0.47
sept,16,1247 LV,0.50,0.46
sept,16,1306 LV,0.50,0.46
sept,16,1247 LV,0.51,0.49
sept,16,1306 LV,0.51,0.49
sept,16,1505 RV,0.50,0.45
```

Header

Note: the columns are only separated with commas.
There is no need for additional spaces between commas.

Newline Characters (\n)

- When we open a CSV file in a program like Excel or an IDE like VSCode, the rows are automatically placed onto their own line for readability.

- However, since computers don’t have eyes, they need to **use a special character called a newline** to know where one line should end and one should begin.

Newline Characters (\n)

- A newline character is represented in Python as '`\n`' and it counts as a single character, i.e., `len('\n') == 1` returns `True`.

```
len('\n') == 1
✓ 0.0s
True
```

- You won't need to worry about this for today's assignment, but (hint) it may be useful for upcoming assignments

```
1  print("***No new line character***")
2  print("Hello World")
3  print("***Include a new line character***")
4  print("Hello\nWorld")
✓ 0.0s
***No new line character***
Hello World
***Include a new line character***
Hello
World
```

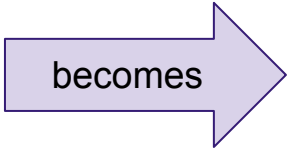
Today's Assignment

- We will be working with a dataset of 2024 US Presidential Election Polls
- The adapted dataset contains the following columns:
  - **month** - month of the year: 'aug' to 'sept' – reminder: they are lower case
  - **date** - day of the month: integer representing the date of the month
  - **sample** - The number of people who responded to the poll and the type of respondents
    - A = Adults
    - V = Voters
    - LV = Likely Voters
    - RV = Registered Voters
  - **Harris result** - the percentage of respondents who preferred Harris
  - **Trump result** - the percentage of respondents who preferred Trump

Your Task

- In the starter code you are given a class `PollReader` and several methods to implement
- `PollReader` reads in the CSV file and builds a dictionary where each key is the name of a column, and each value is a list of the data in that column
- The dictionary is stored in an instance variable called `data_dict`

```
month,date,sample,Harris result,Trump result
sept,19,1880 LV,0.51,0.45
sept,19,1880 LV,0.53,0.47
sept,17,810 LV,0.49,0.45
sept,17,820 RV,0.49,0.45
sept,17,1445 RV,0.49,0.45
sept,17,1000 LV,0.53,0.47
sept,16,1247 LV,0.50,0.46
sept,16,1306 LV,0.50,0.46
sept,16,1247 LV,0.51,0.49
sept,16,1306 LV,0.51,0.49
sept,16,1505 RV,0.50,0.45
```




```
self.data_dict = {
  'month': ['sept', 'sept', 'sept'...],
  'date': [19, 19, 17...],
  'sample': [1880, 1880, 810...],
  'sample type': ['LV', 'LV', 'LV'...],
  'Harris result': [0.51, 0.53, 0.49...],
  'Trump result': [0.45, 0.47, 0.45...]
}
```

Your Task

- First, you'll need to **fix the bugs** in the *build_data_dict()* method
  - HINT: Think about the header row and how columns are separated in CSVs

- Then you'll need to implement each method of the *PollReader* class according to the instructions in the starter code.

- We have provided several test cases for you that should pass if you've completed the assignment successfully

- **Please don't change any of these test cases**

Midterm Review Material

Find the midterm review material in
Canvas > Files > Discussion 6 > Midterm_review.pdf

Remember: You can have a 8.5*11 paper as a cheat sheet for the midterm