Aske Svane Qvist

Exercise W36Wed: Pandas Preview and Data Ethics

1. The SODAS Dataframe

> **Exercise 1.1:** Run the two cells below to read in the SODAS dataset we created. Then, inspect the dataframe and create a markdown cell in which you respond to the following questions:

1. What are the names of the postdoctoral researchers at SODAS?
2. What do you think is the source for the last three columns of the dataframe?
3. In what year was the most highly cited publication in the dataframe published?
4. Three people in the dataframe are each authors on two publications listed in the dataframe. What are their names?
5. Who are the three individuals who have had their Twitter account the longest and since when have they had them?

In [1]:

```
# Loading the pandas module
import pandas as pd
```

In [2]:

```
# Loading the SODAS dataframe

url = 'https://dl.dropboxusercontent.com/s/9war4suj1s5j1ah/sodas_people_twitter_s

sodas_people_df = pd.read_csv(url)
```

In [13]:

```
sodas_people_df.head(3)
```

Out[13]:

	description	role	twitter	google_scholar	
0	David Dreyer Lassen is the Director of SODAS a...	SODAS steering committee	https://twitter.com/daviddlassen	https://scholar.google.dk/citations?user=aRBQc...	david.dreye
1	Morten Axel Pedersen is Deputy Director of SOD...	SODAS steering committee	NaN	https://scholar.google.ca/citations?user=4vDlk...	
2	Rebecca Adler-Nissen is Professor in Political...	SODAS steering committee	https://twitter.com/rebadlernissen?lang=da	https://scholar.google.dk/citations?user=lazTX...	

3 rows × 25 columns

Answer

1. What are the names of the postdoctoral researchers at SODAS?

In [5]:

```
# Create boolean object indicating postdoctoral researchers
postdoc =  sodas_people_df['role'] == 'Postdoctoral Researcher'
# only include 'True'
postdoc = sodas_people_df[postdoc]
# Print the names
print(postdoc['name'])
```

```
13    Hjalmar Alexander Bang Carlsen
14                    Patrice Wangen
15           Kristin Anabel Eggeling
16                  Helene Willadsen
17                   Mette My Madsen
18                       Anna Rogers
Name: name, dtype: object
```

2. What do you think is the source for the last three columns of the dataframe?

Google scholar

3. In what year was the most highly cited publication in the dataframe published?

In [11]:

```
# I find the most cited (boolean object)
most_cited =  sodas_people_df['gs_most_cited_cites'] == sodas_people_df['gs_most_
# Only extract the max
most_cited = sodas_people_df[most_cited]
# Print the year it was published
print(most_cited['gs_most_cited_year'])
```

```
3    2010.0
Name: gs_most_cited_year, dtype: float64
```

4. Three people in the dataframe are each authors on two publications listed in the dataframe. What are their names?

Since the several names have been written together and in different ways, I explore the dataframe manually:

- Sune Lehmann, A Bjerre Nielsen, and R. Adler Nielsen

5. Who are the three individuals who have had their Twitter account the longest and since when have they had them?

In [15]:

```
# Create a new column where the dates are made into a date object
sodas_people_df["twitter_dates"]= pd.to_datetime(sodas_people_df["twitter_createc
```

In [16]:

```
# Sort by dates and print the first three.
sodas_people_df.sort_values(by="twitter_dates").head(3)
```

Out[16]:

	description	role	twitter	google_scholar	
3	Sune Lehmann is a Professor of Complexity and ...	SODAS steering committee	https://twitter.com/suneman	https://scholar.google.com/citations?user=wvkU...	
46	Tobias Priesholm Gårdhus, Master Student in So...	Student Assistant	https://twitter.com/gaardhus	NaN	
11	Gregory Eady is an Assistant Professor in the ...	Assistant Professor	https://twitter.com/GregoryEady	NaN	gregory.ea

3 rows × 26 columns

- Sune Lehmann, Tobias Priesholm Gårdhus, and Gregory Eady

2. Getting Data I - Downloading Data

> **Exercise 2.1:** Browse the data sources we discussed and download one dataset you are interested in. Load it into a pandas DataFrame and inspect the dataframe. Create a markdown cell and describe the following with your group.

1. What information does your DataFrame contain?
2. What is the source of your dataset? Is there a direct link?
3. Are there any striking or particularly interesting datapoints in your DataFrame? What are they?
4. What research question could you answer with this data?
5. Is there any other data you would like to combine with your DataFrame? Which additional data would you like to have?

Some of the data sources we discussed are:

- Inter-university Consortium for Political and Social Research (ICPSR) (https://www.icpsr.umich.edu/icpsrweb/ICPSR/)
- GESIS database (https://search.gesis.org/research_data)
- Kaggle (https://www.kaggle.com/datasets)

Note that all of thise require you to register. If you do not want to register anywhere, Eurostat (https://ec.europa.eu/eurostat/web/main/home) provides a host of statistics about Europe and does not require sign up. For example, here (https://ec.europa.eu/eurostat/estat-navtree-portlet-prod/BulkDownloadListing?file=data/tgs00007.tsv.gz) is a table with statistics on employment rates for age group 15-64 by sex at the regional level across Europe.

Note that, to load your data into a pandas DataFrame, you will have to use the appropriate function for your data format, such as `read_excel` for an Excel file or `read_json` for a JSON file. You can find a list of all the different function and documentation on how to use them here (https://pandas.pydata.org/pandas-docs/stable/reference/io.html).

Answer

```
In [41]: polity = pd.read_excel('p5v2018.xls')
```

In [46]:
```
polity.loc[polity["scode"] == "USA"]
```

Out[46]:

	p5	cyear	ccode	scode	country	year	flag	fragment	democ	autoc	...	interim	bmont
16561	1	21776	2	USA	United States	1776	0	0.0	-77	-77	...	.	
16562	1	21777	2	USA	United States	1777	0	0.0	-77	-77	...	-77	
16563	1	21778	2	USA	United States	1778	0	0.0	-77	-77	...	-77	
16564	1	21779	2	USA	United States	1779	0	0.0	-77	-77	...	-77	
16565	1	21780	2	USA	United States	1780	0	0.0	-77	-77	...	-77	
...	...	...	...	...	...	...	...	...	...	...	...	...	
16801	1	22016	2	USA	United States	2016	0	0.0	8	0	...	8	
16802	1	22017	2	USA	United States	2017	0	0.0	8	0	...	8	
16803	1	22018	2	USA	United States	2018	0	0.0	8	0	...	8	
16804	1	22019	2	USA	United States	2019	0	0.0	7	0	...	7	
16805	1	22020	2	USA	United States	2020	0	0.0	5	0	...	6	

245 rows × 37 columns

What information does your DataFrame contain?

- Countries of the world over a series of time with different scores. The dataframe includes democracy scores, autocracy scores, as well as various other parameters.

What is the source of your dataset? Is there a direct link?

- The data source is 'Center for systematic peace'. The link to retrieve the data is http://www.systemicpeace.org/inscrdata.html (http://www.systemicpeace.org/inscrdata.html)

Are there any striking or particularly interesting datapoints in your DataFrame? What are they?

- The development of the democracy scores of the US throughout the last 6 years. It is striking - but not really surprising.

What research question could you answer with this data?

- In combination with other societal and historical issues, the democracy scores can be used to shed light on the influence of the degree of democracy on other matters.

Is there any other data you would like to combine with your DataFrame? Which additional data would you like to have?

- scores of GDP, the profile of the leading party at any given time.

3. Data Ethics

> **Exercise 3.1:** This activity requires you to familiarize yourself with a case study, discuss a set of questions in your group, and then share and discuss your responses in plenum.

Tasks and Suggested Schedule

1. Take *10 minutes* to read this case study (https://bdes.datasociety.net/wp-content/uploads/2016/10/Patreon-Case-Study.pdf) describing a data ethics issue encountered by two social data scientists.
2. Discuss these four questions in your group for *20 minutes* and take notes about your answers:
    - Does the illegal nature of the data collection and the release of private data taint the data in the release that was already publicly available?
      you are not making anything more public. the harm is already done do one could argue that it is in the pulbic interest to optimize benefice
    - Users of Patreon initially had an expectation of privacy, but that privacy no longer exists. Do researchers need to respect the intent or the reality?
    - Researchers will nearly always claim that their research will have a net public benefit and thus their methods are justified. Who gets to decide if that is accurate in any given case?
    - Would you personally use the Patreon data in this situation?
3. Reconvene as a large group and share the answers your group came up with. Do you all agree? On what do you disagree and why? *Approximately 15 minutes.*

Notes

- Talking about ethics can be difficult for some people. Please keep an open mind, and remember that reasonable people can disagree on many of these topics.
- This exercise is adopted from material developed by Matthew Salganik, Robin Lee, Ian Lundberg, Yo-Yo Chen, Janet Xu, and Simone Zhang for the Summer Institutes in Computational Social Science (https://compsocialscience.github.io/summer-institute/)

3. Creating the SDS Imagination

> **Exercise 4.1** A key ingredient for good Social Data Science is awareness of the data out there, what others are doing, and what others find interesting. This excercise is intended to provide a venue to build this awareness.

For each group:

1. Present the data you collected in 2. Getting Data I - Downloading Data to the larger group touching on all five questions. *Approximately 5 minutes.*
2. Discuss the following prompts in the larger group: *Approximately 5 minutes.*

    - Is the dataset suited to the proposed research questions? What are the strengths of the data for this purpose?
    - What are potential shortcomings of the data? Are there alternative data sources without these shortcomings? What are they?
    - Which *other* research questions could be answered with the dataset? Would additional data be necessary? Which?

3. Repeat for each group. Take notes when your group is *not* presenting.

> **Exercise 4.2:** As you know, sharing is caring. Share the link to the dataset you found and a little description of the type of data it contains in this shared document (https://docs.google.com/spreadsheets/d/1-SOaU7QHXWC9lieYVMVi_xEOUn3zXlm11UwLBpuBAiI/edit?usp=sharing). Your fellow SDS classmates will appreciate it.

5. Daily Reflections

> **Exercise 5.1** Take a moment to reflect on your learning experience today and take notes. You *can* use these prompts to inspire your reflections:

What (if anything) did you take away from the lecture and exercise today?

- It was nice to see how easy it can be to find so much information online. The different groups had found so many interesting things that can be used to answer all sorts of questions. I am excited!

What concepts, ideas, or topics are still unclear?

- I think the discussion of ethics is still super complicated to grasp - and will probably continue to be so.

Are there any things you would have wanted to spend less or more time on? What are they?

- I am really excited about starting to code more. But I guess I will be more than stimulated soon enough

6. *If You Have Time* Follow Your Interests

> **Exercise 6.1** Think about the different data sources and datasets you have heard about in 3. Creating the SDS Imagination. Follow up to see if you can track down one of the alternative data sources discussed or find data relevant to one of the other research questions that seemed interesting to you.

If you are successful, load the data into a pandas DataFrame and inspect them. If you do not find relevant data, take notes on what you tried, where you looked, and follow up with your group members to see if they have additional leads.

Remember that there is an entire Dataverse (https://dataverse.harvard.edu/) out there and recently even Google has stepped into the game of cataloging publicly available data (https://datasetsearch.research.google.com/).