Who Runs the World?

An analysis of women in world governing bodies by Samantha Roska, Rebecca Hailperin-Lausch, and Samantha Russel

Motivation

It has been shown that more women in government can lead to benefits for all citizens, such as increased spending on health and education, even leading to decreases in a country's mortality rates (Ng and Muntaner). The 118th United States Congress boasted the highest percentage of women in history, with women making up 28% of the governing body (pew research). This fact may seem to champion the US as a model of equality, however, how does this claim hold up when compared to other countries? Using data from the *Inter-Parliamentary Union and the Economist Intelligence Unit* democracy index, we will explore how women's leadership in government has changed in recent years. In addition to comparing gender division percentages between the governing bodies of democratic countries, we took the question one step farther to look at how the number of women in government compares to democracy scores for each country: do countries with a higher percentage of women in government boast higher democracy scores? And, knowing that different areas of the world have different cultural expectations when it comes to women, how do these relationships change when examined by region?

Data Sources

Data Source 1: Women in Parliaments

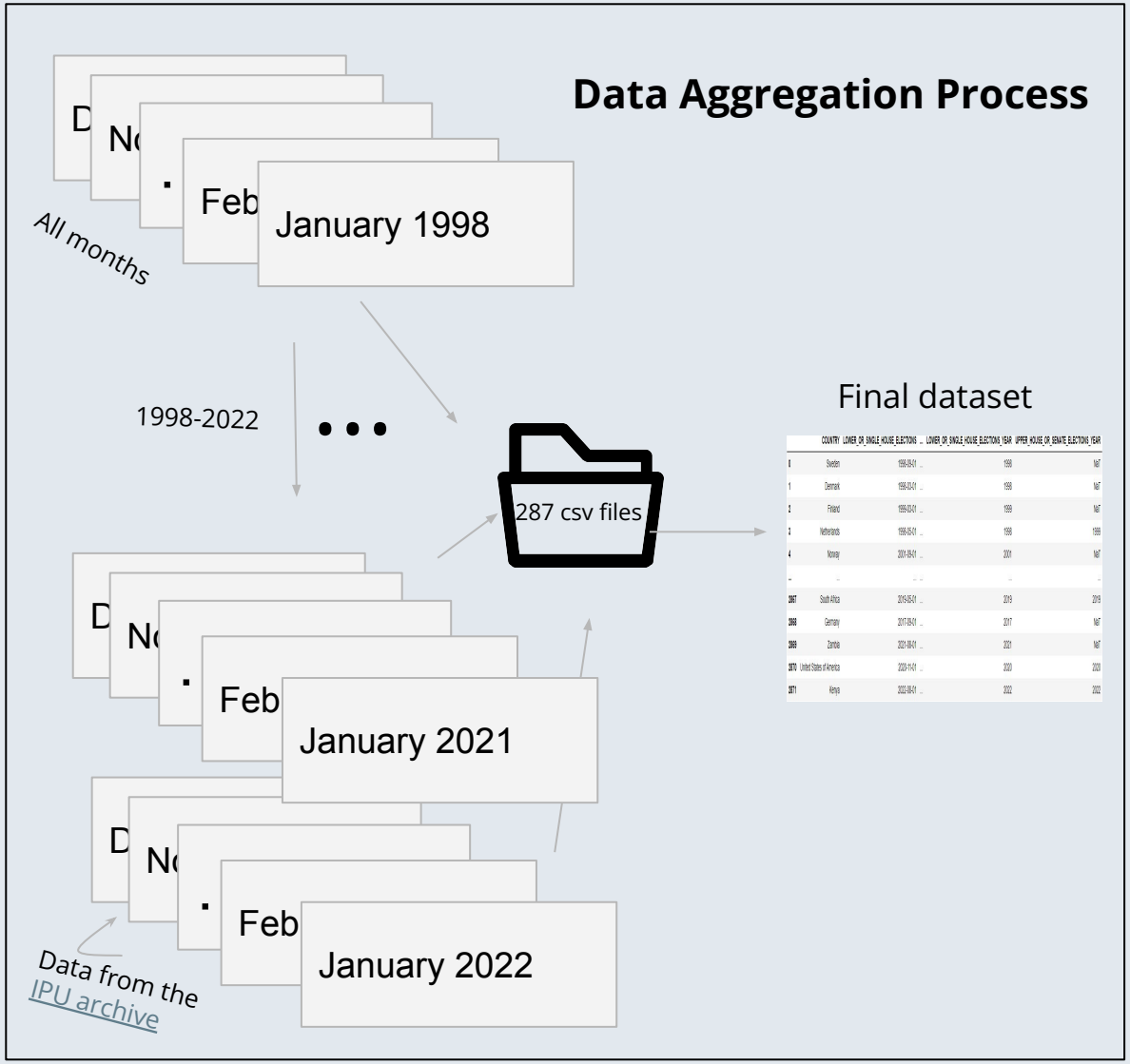
This dataset comes from the Inter-Parliamentary Union (IPU) and contains information about the number of women in parliaments across the world, with data from 1998 to 2022 available.

Important Variables

- country
- lower_single_house_elections (election month and year for lower house)
- lower_single_house_seats (total number of seats in lower house)
- lower_single_house_women (number of women in lower house)
- lower_single_house_percent_w (percent women in lower house)
- upper_house_senate_elections (election month and year for upper house)
- upper_house_senate_seats (total number of seats in upper house)
- upper_house_senate_women (number of women in upper house)
- upper_house_senate_percent_w (percent women in upper house)

This dataset was obtained using a python script which scraped the data from the IPU archive for each month of each year into individual csv files. These files were then aggregated together to create a main csv file that contains the data with the percent women in parliaments across the world from 1998 to 2022.

The raw csv files totaled 267 MB. After aggregation, the main dataset contains 9833 rows.



This diagram shows the data aggregation process for the women in parliaments dataset. The final dataframe has 9833 rows and 12 columns, with data from 1998-2022.

Data Sources

Data Source 2: Democracy Index

The Democracy Index is an indicator produced by the Economist Intelligence Unit to measure the state of Democracy in countries around the world. The scores range from 0-10, with a higher score indicating a country is more democratic and a lower score indicating a country is more authoritarian. The democracy index scores were available for most years from 2006-2022.

Important Variables

- Region
- Country
- Years (2006-2022)
- Democracy Index Score (0-10)

This dataset was obtained from wikipedia and was read in directly using pd.read_html. This was then saved to a csv file.

This file is 23 kilobytes and contains 167 records.

	Region	2022 rank	Country	Regime type	...	2010	2008	2006
0	North America	12	Canada	Full democracy	...	9.08	9.07	9.07
1	North America	30	United States	Flawed democracy	...	8.18	8.22	8.22
2	Western Europe	20	Austria	Full democracy	...	8.49	8.49	8.69
3	Western Europe	36	Belgium	Flawed democracy	...	8.05	8.16	8.15
4	Western Europe	37	Cyprus	Flawed democracy	...	7.29	7.70	7.60
...	...	...	...	...	...	...	...	...
162	Sub-Saharan Africa	92	Tanzania	Hybrid regime	...	5.64	5.28	5.18
163	Sub-Saharan Africa	130	Togo	Authoritarian	...	3.45	2.43	1.75
164	Sub-Saharan Africa	99	Uganda	Hybrid regime	...	5.05	5.03	5.14
165	Sub-Saharan Africa	78	Zambia	Hybrid regime	...	5.68	5.25	5.25
166	Sub-Saharan Africa	132	Zimbabwe	Authoritarian	...	2.64	2.53	2.62

This image shows a portion of the raw democracy index dataframe. The full dataframe had 167 rows and 20 columns, with data from 2006-2022. This dataset had no missing values, although data from 2007 and 2009 were not available.

Data Manipulation Methods

Pre-processing

The largest dataset is *Women in Parliament*. This data spans **287 .csv files** with different header formats. Each file required reading (*pandas.**read_csv**)* and a programmatic check for discerning the appropriate first row of values. After processing each .csv file into a single dataframe, the resulting datafarme was placed into an array until all 286 files were read. The array of 286 dataframes was then merged (pandas.**merge**) into one dataframe with **~52,863 rows**.

All the source files were consistent with the number of columns, this simplified the processing, and only required renaming (pandas.**rename**) the column headers.

Among the 286 files were **3 different header formats**. It took some manual review of the files to notice that the header format followed a pattern. File names were suffixed by year and followed the following pattern:

- Years after 2019 had **5** header rows
- Years 2008 - 2019 had **2** header rows
- Years before 2008 had **1** header row

Discovering the header pattern allowed for a **regex** check on the file name and then skipping the correct amount of header rows to guarantee the first row of the dataframes were numerical values across 11 columns.

Calculation Considerations

Nearly all the data sources used included raw counts of women in public office positions. In order to get statistical insights the data required *Pandas* functions for calculating percentage, mean, and sum over time. However, the source data was either not in a numeric format or contained a character of some kind (“-”,” “).

Columns with a numeric value not of the type (pandas.**dtype**) numeric, were converted using the pandas.**to_numeric** function. There was only a need for checking fields, using ***regex***, for a character that was not a number. If the regex found such a character, it was removed from the field. **For example: ‘2.3%W’ → 2.3**

After columns were converted to numeric format that left some fields with NaN (not-a-number). On first thought, it could be assumed to drop the NaN values. However, when calculating sum or average it is much better to convert the NaN to 0 (zero). We did not want to skew any analysis because of valuable data that was dropped.

Numeric data that was representing a percentage but was originally encoded as string had to be converted. The conversion was not the data-type (it had already been converted to numeric), rather it was the value itself was not correct. the **to_numeric function was translating 2.3 into 2300 within the dataframe**. This was rectified by replacing the provided percentage values with calculated percentages from the data. Example code snippet:

```
dataframe.percentage_field = dataframe.women_in_seats / dataframe.total_num_seats
```

Data Manipulation Methods : Dataframes

Naming Conventions

Among all data sources used for analysis the column naming and conventions were not consistent. While not uncommon, having varied naming can result in duplicate code and inconsistent referencences within analysis.

We established a naming convention for **columns names to follow snake_case**; lowercase words separated by underscore. Using pandas.**rename.** This added consistency and cohesion among the code we wrote individually.

Additionally, for the column names we had to pandas.**replace** non-alpha characters like '-' with an empty the value. Effectively removing the character.

Country names are central to our analysis and visualizations. In order to be consistent within our code **we made a file (*/data/list_of_countries.csv*) containing the spelling conventions for country's**. For example: 'United States' in place of 'United States of America'.

Grouping Series Data

The heart of our analysis is based on correlations to answer our main question - Do countries with a higher democracy index have more women in their government? Our initial exploration of correlation between Index Score, Lower House %Women, and Upper House %Women yielded a weak-positive correlation. This lead us to question whether region played a more important role in the relationship. Performing a pandas.**groupby** on 'region' afforded us the ability to confirm our hypothesis. There were differences in correlation by region for values in Index Score, Lower House %Women, and Upper House %Women. More detail on this result in provided on slide 10.

Dropping Duplicate Data

Within *The Women in Parliament* dataset each .csv has a column for the date of the country's Election (mm/yyyy). Given that not all countries hold their elections on the same month within a year, there were duplicate rows within the combined dataset (~52k rows). To handle this issue we **first converted all the column's data types** appropriately:

```
country                          object
lower_single_house_elections
datetime64[ns]
lower_single_house_seats          float64
lower_single_house_women          float64
lower_single_house_percent_w      float64
upper_house_senate_elections
datetime64[ns]
upper_house_senate_seats          float64
upper_house_senate_women          float64
upper_house_senate_percent_w      float64
```

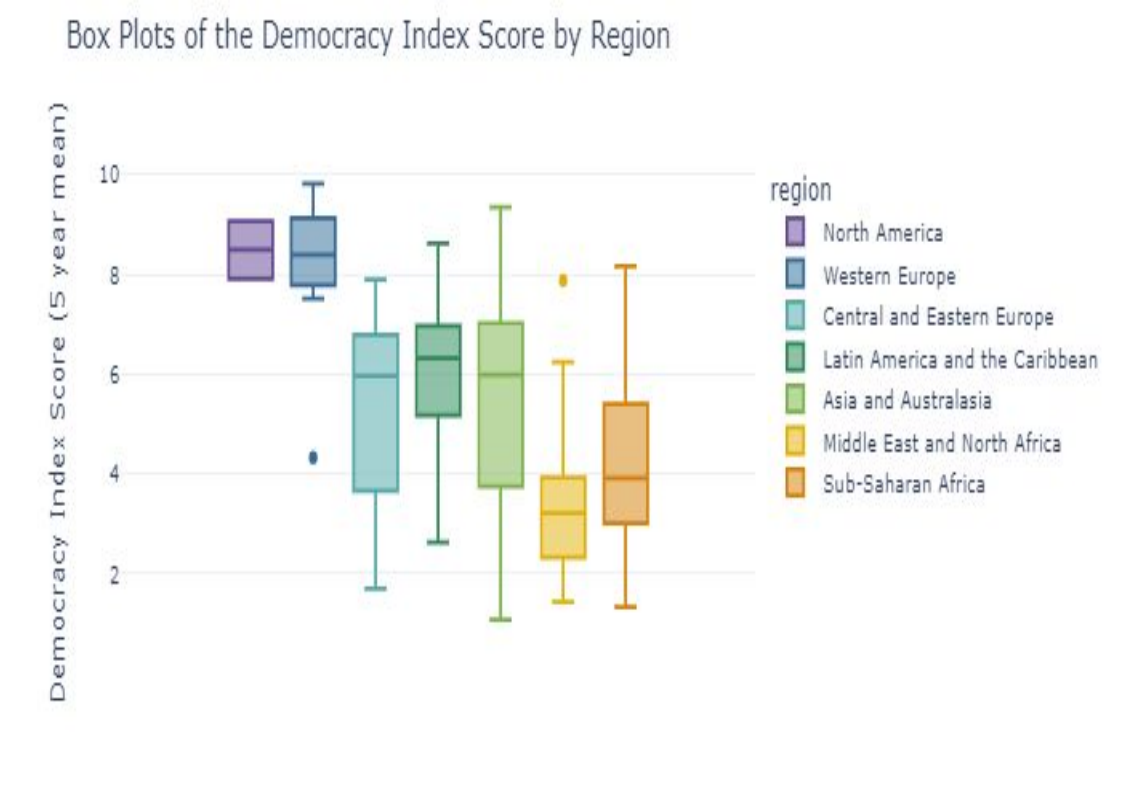
and **then used pandas.drop_duplicates** to remove the rows that were exactly the same across all the dataframe's columns. This changed the number of rows from ~52k to ~10k.

Merging Dataframes

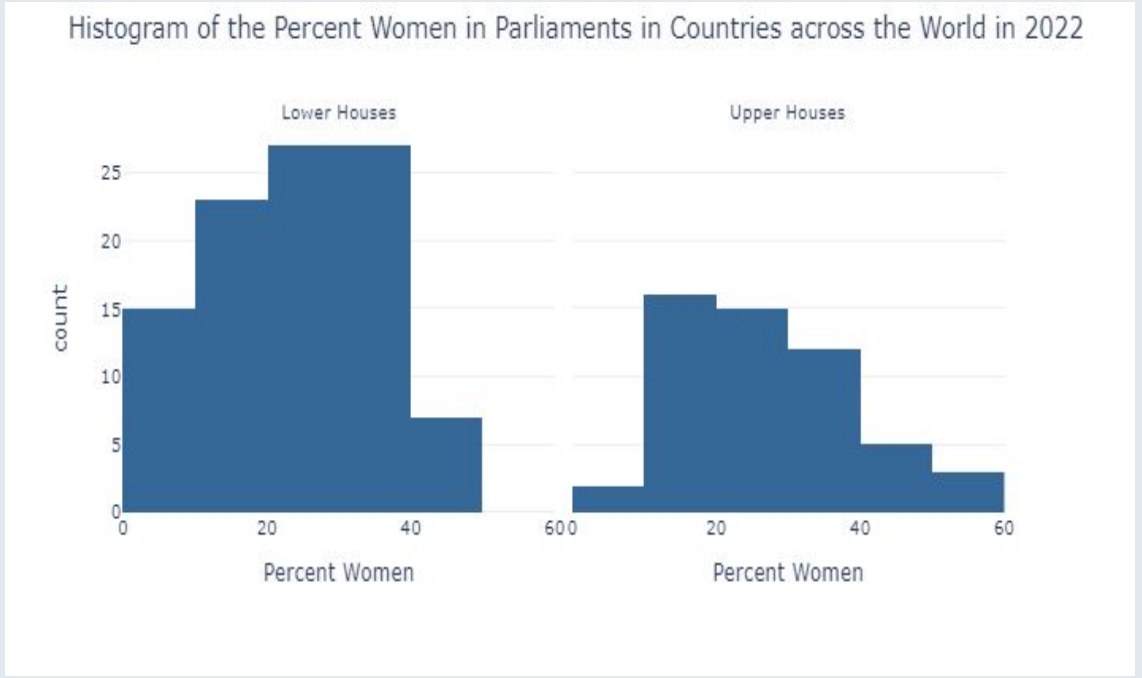
Both primary data sources are categorized by a Country name, a string value, that was used to join dataframes. In order to join, without losing data, this required **specifying the join type as "left"** because an "inner" join would only return rows with an exact match on a string. By doing a "left" join we maintained data integrity and could easily find rows with Country name conflicts. Conflicting rows result in "NaN" for the field being joined on. Using **pandas.isna** on 'Country' we filtered the dataframe for the value "True" in order to **individually manipulate the rows that had different Country spelling between the dataframes.**

**Note that in some data manipulations the Country's ISO code (string value) was used as the join condition. This change of data field did not change the manipulation process mentioned above.*

Analysis



This visualization shows the box plots of the democracy index scores grouped by region. Democracy index scores range from 0-10. Western Europe and North America show the highest mean democracy index scores and smallest range, while the Middle East show the lowest democracy index scores.



This visualization shows histograms of the percent women in parliaments in countries across the world in 2022. We saw considerable differences between the percentages of women in upper vs. lower houses. Lower houses had more countries with higher percentages of women in their parliaments, with the majority of countries falling between 20-40%. On the other hand, upper houses showed lower percentages of women overall, but did have a higher max, with three countries having between 50-60% women in their upper house.

Analysis

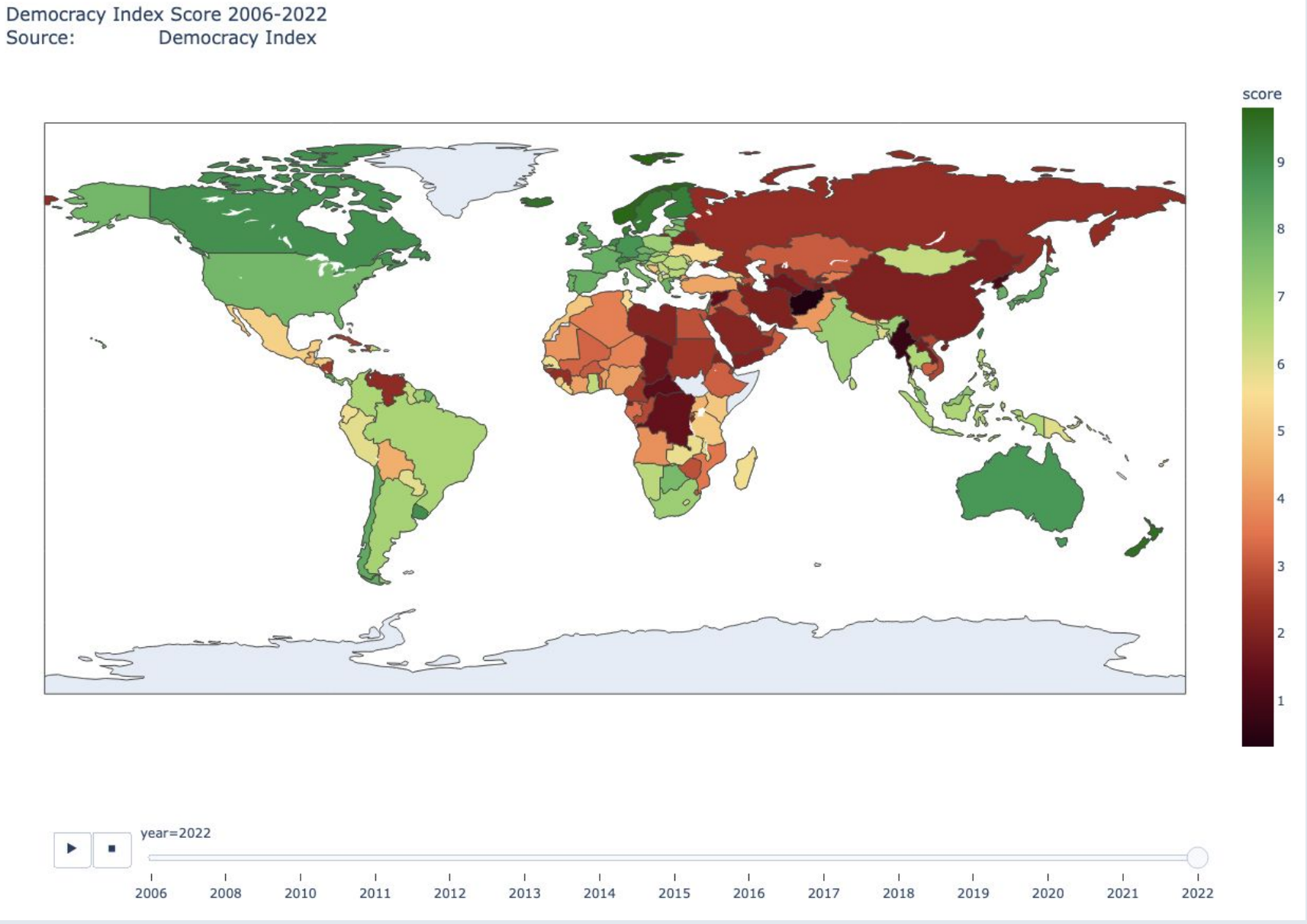
We found that democracy index scores varied broadly across the world. In the animated version, you can see that most countries' scores do not very much from 2006 to 2022. Given the large variation in democracy index scores, we wondered if this variable would in any way be correlated to the large variation in the percent of women in government we also saw across the world.

Full democracies
- 9.01–10.00
- 8.01–9.00

Flawed democracies
- 7.01–8.00
- 6.01–7.00

Hybrid regimes
- 5.01–6.00
- 4.01–5.00

Authoritarian regimes
- 3.01–4.00
- 2.01–3.00
- 1.01–2.00
- 0.00–1.00

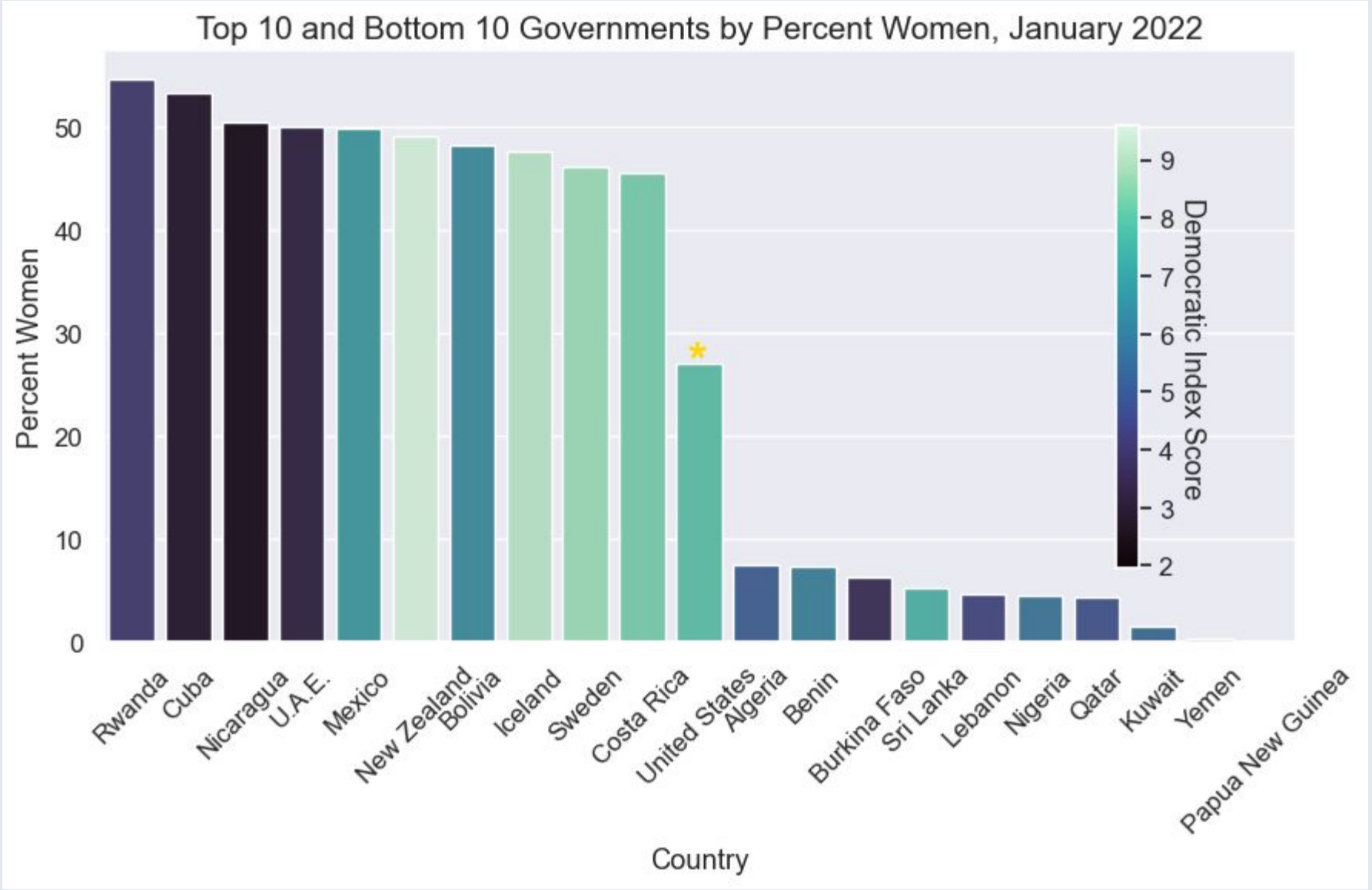


https://github.com/samroska/women-in-gov/blob/main/data_visualizations.ipynb

Analysis

Bar Chart Comparing the Percent of Women in Government Across Countries, Colored by Democracy Index Scores

This visualization shows the top 10 and bottom 10 countries in terms of the percent of their governing bodies that are women. We also included the United States for reference. We found that the countries with the most women did not necessarily boast the highest democracy index scores, and vice versa. Given that there was not a noticeable trend at the world level, we decided to explore this relationship by region to account for possible regional cultural differences.



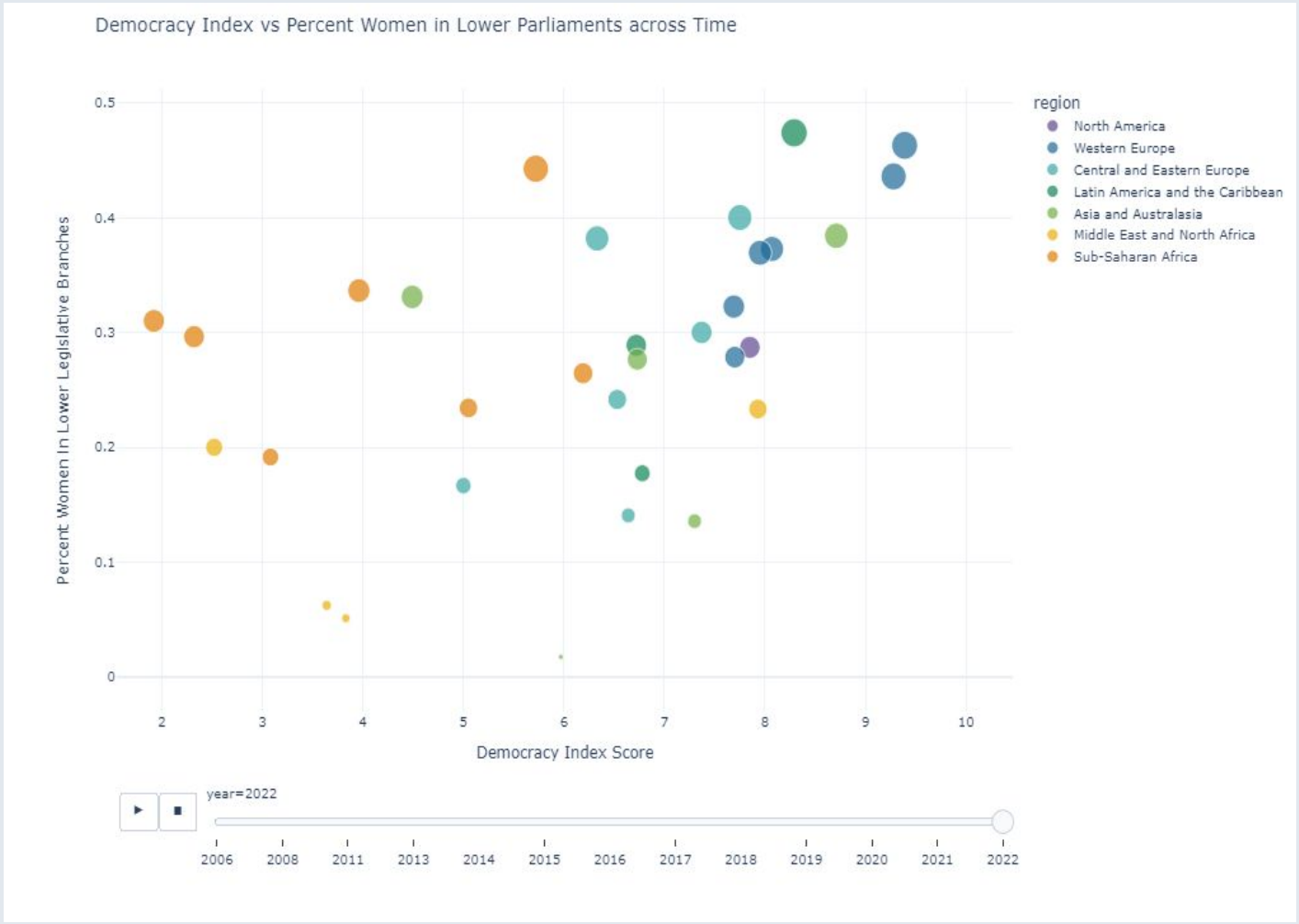
Analysis

Scatter Plot of Democracy Index Scores vs Percent Women in Parliament

This visualization shows the democracy index score vs. percent of women in lower legislative branches. We decided to focus on the lower houses only for this visualization because upper houses showed a weaker correlation, and because upper house elections tend to happen less frequently, leading to less data being available. Each point in the plot represents a country. The color of the points indicate the region and the size of the points help to emphasize the percent of women. This visualization shows the data from 2022, but the interactive visualization can be found in the notebook and as an html file in the project folder.

There is a slight, upward trend in this scatterplot. This makes sense given that we found that the democracy index score and percent women in lower houses have a moderate, positive correlation of 0.31 (see next page for more details).

By using color to encode region, it becomes evident that there are clear differences by region, mostly along the x-axis. The left half of the scatter plot is mostly yellow and oranges (Middle East Africa), while the right half of the plot is most blues and greens (Europe, America's and Asia). This finding mirrors what we found with the box plots on pg 6 of this report.



Analysis

Correlation matrices

The top correlation matrix shows the correlations between the democracy index score and the percent women in upper and lower houses. Unsurprisingly, lower and upper houses showed a moderately strong, positive correlation of 0.63. This suggests that countries with a higher percent of women in lower houses, tend to also have a higher percent of women in their upper houses and vice versa.

The democracy index score and percent women in lower houses have a moderate, positive correlation of 0.31, while the the democracy index score and percent women in upper houses showed a weak correlation of 0.18. Because of these findings,, we decided to look at the correlations between these variables by region.

The correlations by region showed no consistent pattern, with some regions showing negative correlations and some regions showing weak or positive correlations. This could help to explain why the overall correlations were moderate to weak. After looking into it further, it seems likely that this wide variation is partially due to the small number of data points within in each region.

	democracy_index_score	lower_single_house_percent_w	upper_house_senate_percent_w
democracy_index_score	1.000000	0.310565	0.180674
lower_single_house_percent_w	0.310565	1.000000	0.626076
upper_house_senate_percent_w	0.180674	0.626076	1.000000

region		democracy_index_score	lower_single_house_percent_w	upper_house_senate_percent_w
Asia and Australasia	**democracy_index_score**	1.000000	0.128786	0.293848
	lower_single_house_percent_w	0.128786	1.000000	0.885088
	upper_house_senate_percent_w	0.293848	0.885088	1.000000
Central and Eastern Europe	**democracy_index_score**	1.000000	0.067009	-0.442193
	lower_single_house_percent_w	0.067009	1.000000	0.422929
	upper_house_senate_percent_w	-0.442193	0.422929	1.000000
Latin America and the Caribbean	**democracy_index_score**	1.000000	-0.217821	0.327432
	lower_single_house_percent_w	-0.217821	1.000000	0.824936
	upper_house_senate_percent_w	0.327432	0.824936	1.000000
Middle East and North Africa	**democracy_index_score**	1.000000	0.376117	-0.500286
	lower_single_house_percent_w	0.376117	1.000000	0.201189
	upper_house_senate_percent_w	-0.500286	0.201189	1.000000
North America	**democracy_index_score**	1.000000	0.205076	-0.971636
	lower_single_house_percent_w	0.205076	1.000000	0.914636
	upper_house_senate_percent_w	-0.971636	0.914636	1.000000
Sub-Saharan Africa	**democracy_index_score**	1.000000	0.113238	-0.066918
	lower_single_house_percent_w	0.113238	1.000000	0.726346
	upper_house_senate_percent_w	-0.066918	0.726346	1.000000
Western Europe	**democracy_index_score**	1.000000	0.545194	-0.191058
	lower_single_house_percent_w	0.545194	1.000000	0.572877
	upper_house_senate_percent_w	-0.191058	0.572877	1.000000

Conclusion

While progress certainly has been made, the US is by no means the model for women equality in government. We found they ranked fairly middle of the road in their percent of women in governing bodies, and also found that the US does not have as high of a democratic index score as we previously thought. We also found the while women in government and democracy scores are positively correlated, the relationship at the world level is moderately weak. We thought that this could be partly due to regional cultural differences and expectations for women. However, when we examined this relationship by region, the results were only muddied further. Some regions still had this moderate-weak correlation, while others actually saw negative correlations. Breaking down the correlations into this small of subsets may have rendered them too small for meaningful analysis. Further research could be done into what specific characteristics about each region contributed to this finding and what other possible confounding variables could be at play. It also remains a possibility that democracy index scores and the number of women in government for countries do not have a relationship, which is in itself an interesting finding.

Statement of Work

The history of our github commits can be found here.

Samantha Roska: Github setup and initialization, democracy choropleth, data manipulation section

Rebecca Hailperin-Lausch: Data scraping & collection, scatterplot, data description section, boxplot and histogram, correlation matrices

Samantha Russel: Team organization and communications, bar charts, motivation and conclusion section

All: Data cleaning and joining, regular team meeting attendance, editing final report

Notes for Future Work: We all feel we collaborated very well together. However, there was an opportunity for growth to do differently in the future to enhance our workflow. Even though we used GItHub, we all started work in different Jupyter notebooks. This forced us to have a multi-day work session later on to combine our files into one notebook. In the future, we would instead work on one file at different times, pushing our individual changes after each session.