Part I: Data Work Descriptor (Materials and Workflow Documentation Provided in Github)

1.1 Learning Objectives

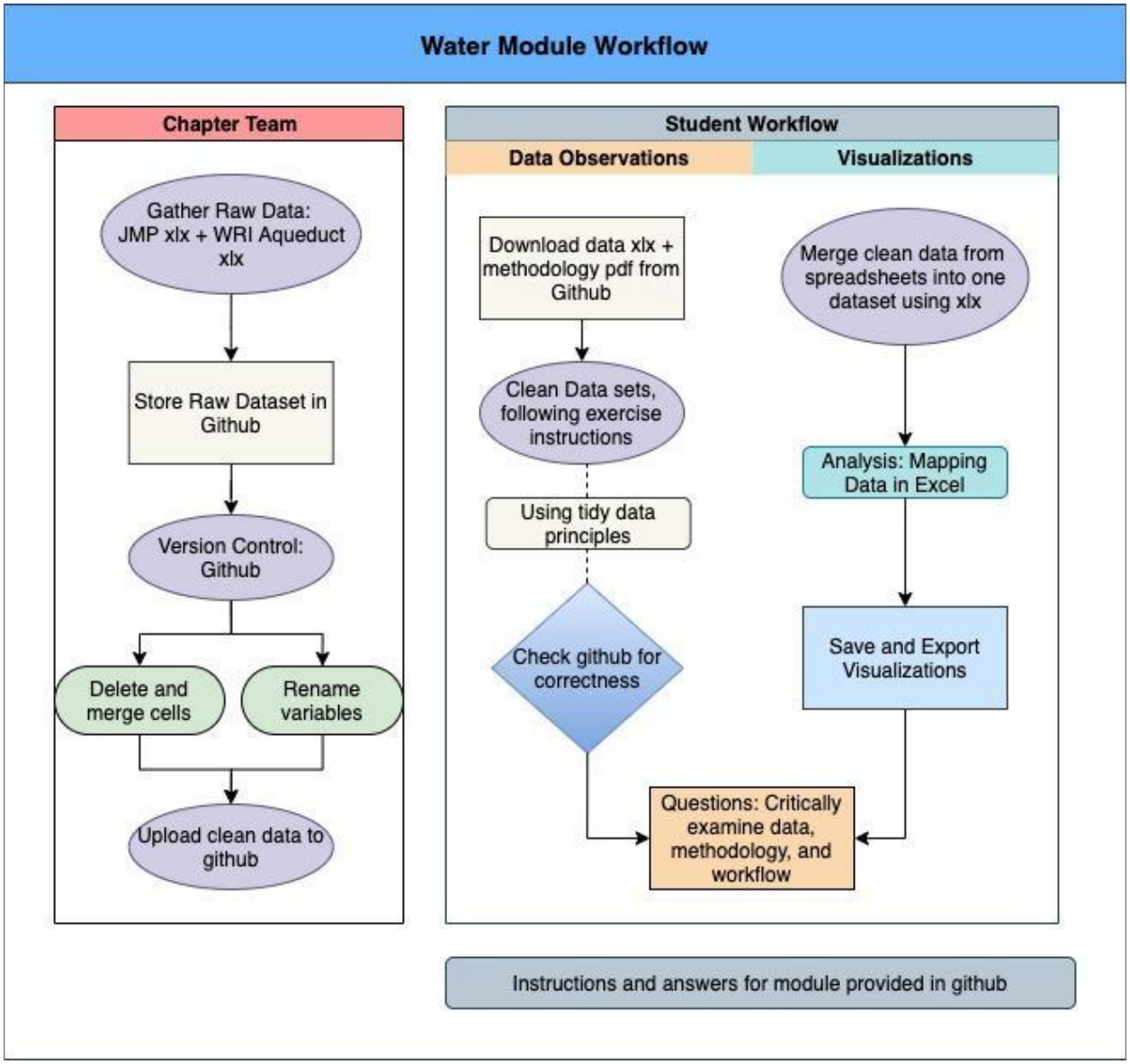
The purpose of this data module is to provide students with the analytical tools to explore various water datasets. We will explore how to conceptualize, research, and visualize common water indicators. In doing so, our mission is to uncover how these indicators can be misused or misinterpreted. The module will be split into two sections. Section one will focus on exploring WHO / UNICEF Joint Monitoring Programme for Water Supply, Sanitation and Hygiene (JMP) and World Resources Institute (WRI) Aqueduct data and methodology. Students will learn how to find, download, and read the data and methodology. We will ask critical questions surrounding the metadata and guide participants to use the tools and analyses provided in Chapter 10's manuscript to answer corresponding questions. Section two of the module will teach students to produce their own visualizations in excel. These visualizations will include research level maps illustrating water access and stress. This data module will encourage students to reflect on biases, capacity, and accountability challenges discussed in section four of the chapter.

Monitoring and evaluating progress on water coverage is paramount to any strong development plan. Importantly, this module will teach students the principles of good data management. We will implement tidy data principles such as: 1) storing all data in a repository, 2) manipulating data using copy files, 3) and ensuring that variable names are clearly defined and understood. In addition, our workflow and github repository allows students to replicate our analysis. Reproducibility is important for developing coherent data practices and allowing for outside sources to reproduce and check the measures. Reproducibility also encourages students to draw their own conclusions when conducting this analysis.

One of the major challenges we encountered when creating this data module was finding comparable data to use. To map water access at a global scale we needed data comprehensive country level datasets. While there are many water datasets the comparability is not always present. Some data provides high level information on regional water issues, others datasets have a lot of missing information. In addition, finding a constructive way to illustrate the politics within water data from only two sampled datasets proved difficult. It is our intent to illustrate the

politics behind water data, not solely the politics within datasets. Therefore, in constructing this module we were cognizant towards teaching students how data can be manipulated for political purposes in combination with showing how the datasets used can be political in nature.

1.2 Data Workflow Chart



1.3 Detailed explanations of all the steps in the workflow.

<u>Data Management</u>

The workflow above is broken up into two parts, part one outlines the work that the chapter team members have done to prepare this data module. Importantly, we begin by gathering all the raw datasets. The two datasets used in this module include: the JMP 2017 WASH data and WRI 2019 Aqueduct 3.0 Country Rankings dataset. The JMP provides data on global water, sanitation, and hygiene coverage. For the purposes of this module, we will focus solely on water coverage. While the WRI dataset provides rankings of country and province water stress levels. The WRI provides rankings based one exposure to overall water risks (see WRI Methodology Aqueduct Technical Note 3.0).

We have downloaded and stored the raw data versions and methodology of the JMP and WRI data in our github repository. All revisions to these data files can also be found in the github repository. Importantly, we also instruct students to make copies of their raw data files and store them in a safe location. These processes follow the data management practices outlined in Wilson et al. 2017 paper *Good Enough Practices in Scientific Computing*.[1] Importantly, taking the proper steps to store is essential for documenting our analysis process and creating a seamless and reproducible workflow.

Software and Version Control

Once the raw data were stored, we began the data cleaning process. All version control is done through excel software and github. We have decided to use Microsoft Excel to complete this data module since this software is easily to learn and highly accessible. This will give students the opportunity to create simple and detailed visualizations through a commonly used software. The repository specifically provides updated excel files highlighting the changes we have made throughout the data module. Throughout the data exercise students are welcome to check the excel files in github to ensure they are doing the work correctly.

Messy data is considerably hard to read and visualize, therefore we include tidy data principles for cleaning the datasets. The members of this chapter have performed the same data cleaning process we ask the students to do within the exercises. This process mainly consists of formatting headers, ensuring that the naming convention for datasets are clear and concise, merging workbooks, and removing unnecessary data. The purpose of following clear formatting

[1] Wilson, Greg, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K. Teal. "Good Enough Practices in Scientific Computing." *PLOS Computational Biology* 13, no. 6 (June 22, 2017): e1005510. https://doi.org/10.1371/journal.pcbi.1005510.

rules is to structure the datasets to improve the analysis process.[2] To ensure the visualizations produced in part II follow these tidy data principles color and shade our maps. This is important to ensure that the data we produce is analysis friendly and easy to interpret.

More details on the steps and the questions we will ask of students can be found in part II. We also include a README file which provides detailed descriptions of the data, github organization, and module exercises. In github students will also find the full instructions for completing the data module.

Collaboration

Overall communication between chapter collaborators this data exercise was executed through text, github, and google documents.

Part II: Data Exercise (Student/Instructor-facing materials provided in the online data modules linked to the printed textbook)

Link to github module materials and osf.

2.1 Learning objectives

The purpose of this data module is for students to familiarize themselves with two different international water datasets, the WHO / UNICEF JMP data on drinking water access and the World Resource Institute (WRI) Aqueduct 3.0 Country Rankings. Exercise one will focus on the differences between JMP and WRI water indicators, measures, and methodologies. In addition, students will learn tidy data principles through cleaning and merging the datasets. The second exercise of this module focuses on developing visualizations for both datasets.

These exercises will focus on JMP water coverage data which is used to measure SDG 6.1.1 - the proportion of population using *safely managed* drinking water services. The WRI dataset we will be using focuses on water risk based on baseline water stress, riverine flood risk and drought risk. These two datasets both tell different stories about water coverage. This module

[2] "Tidy Data." Accessed May 11, 2021. https://cran.r-project.org/web/packages/tidyr/vignettes/tidy-data.html.

will have students evaluate these differences and understand how the same concepts, mainly water access or lack of water access, can be explained using different data and result in different conclusions. We will also discuss the questions that arise when evaluating the datasets and the indicators that may be missing.

By the end of this module, students will be able to access a variety of water data from a critical lens, and know how to extract, clean, visualize, and interpret data collection processes. The exercise questions within the module are meant to guide students on the questions they should consider when reviewing any dataset.

2.2 Exercises 1 and 2

Exercise #1: JMP and WRI Data Comparison

1. Visit Chapter 10's gitHub repository titled Water Module
2. Go to RawData and download the JMP (JMP_2019_WLD.xlsx) and Aqueduct (aqueduct-30-country-rankings.xlsx) excel datasets
3. Return to github and click the Metadata Folder
4. Download the JMP Methodology (JMP-methodology-2017-update.pdf) and the Aqueduct 3.0 Technical Note (aqueduct-30-Technical Note.pdf)
5. Now that the data is downloaded, open both the excel data and methodology report on your desktop
    a. Take a few minutes to familiarize yourself with the both datasets, paying close attention to what they measure and what we can infer from the data.

<u>Part 1 Exploring the JMP methodology and data</u>

6. The first part of this exercise will focus on JMP data. Open the methodology pdf (JMP-methodology-2017-update.pdf)

JMP METHODOLOGY
2017 UPDATE & SDG BASELINES

March 2018

7. Take a few moments to skim the document, paying close attention to how water data is collected and measured. Table 1 and Table 3 (on pages 4 and 9) provide great summaries of the classifications of water indicators.
8. Now go to section 3 Producing estimates (page 13) and read over the information on the primary indications (section 3.2).

Exercise #1 - Please answer the following questions

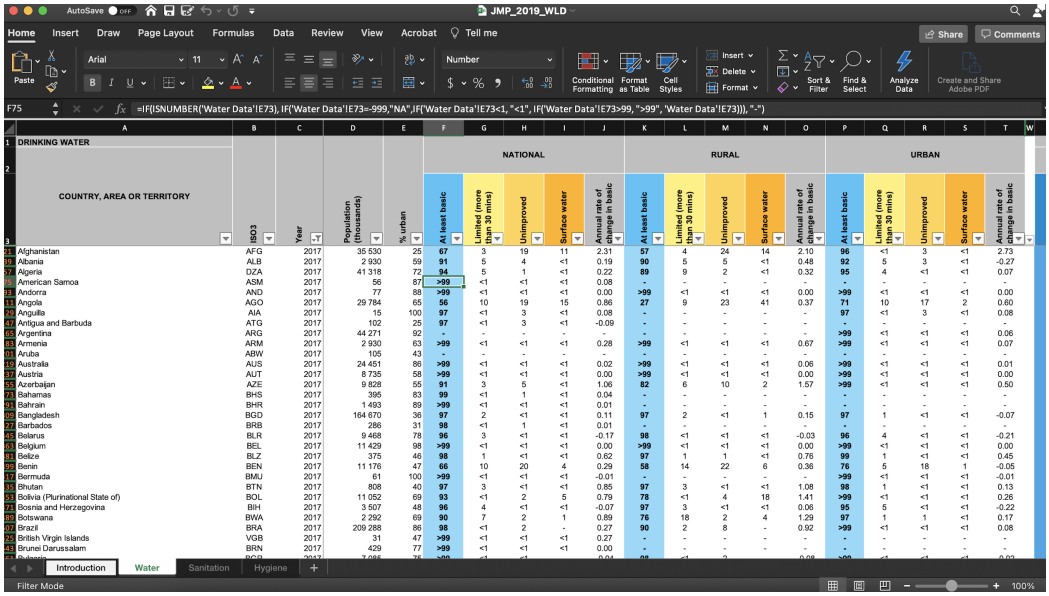
Question 1: What is the JMP and what does it measure?

Question 2: How does the JMP collect water data for each country?

Question 3: What does improved drinking water sources mean?

Question 4: In thinking about how we measure access to water, what are some of the elements that these indicators don't take into consideration? Where could measurements be improved?

9. Now we will explore the JMP dataset (JMP_2019_WLD.xlsx) There are four tabs entitled Introduction, Water, Sanitation, and Hygiene. For this exercise we will be focusing on the Water tab. At the top of the Water tab, you will notice the data is organized by column and provides information on the country (area or territory), year of data observations, population, and dimensions of water access.



JMP_2019_WLD

F75 =IF(ISNUMBER('Water Data'!E73), IF('Water Data'!E73=-999,"NA",IF('Water Data'!E73<1, "<1", IF('Water Data'!E73>99, ">99", 'Water Data'!E73))), "-")

DRINKING WATER

COUNTRY, AREA OR TERRITORY	ISO3	Year	Population (thousands)	% urban	NATIONAL					RURAL					URBAN				
					At least basic	Limited (more than 30 mins)	Unimproved	Surface water	Annual rate of change in basic	At least basic	Limited (more than 30 mins)	Unimproved	Surface water	Annual rate of change in basic	At least basic	Limited (more than 30 mins)	Unimproved	Surface water	Annual rate of change in basic
Afghanistan	AFG	2017	35 530	25	67	3	19	11	2.31	57	4	24	14	2.10	96	<1	3	<1	2.73
Albania	ALB	2017	2 930	59	91	5	4	<1	0.19	90	5	5	<1	0.48	92	5	3	<1	-0.27
Algeria	DZA	2017	41 318	72	94	5	1	<1	0.22	89	9	2	<1	0.32	95	4	<1	<1	0.07
American Samoa	ASM	2017	56	87	>99	<1	<1	<1	0.08	-	-	-	-	-	-	-	-	-	-
Andorra	AND	2017	77	88	>99	<1	<1	<1	0.00	>99	<1	<1	<1	0.00	>99	<1	<1	<1	0.00
Angola	AGO	2017	29 784	65	56	10	19	15	0.86	27	9	23	41	0.37	71	10	17	2	0.60
Anguilla	AIA	2017	15	100	97	<1	3	<1	0.08	-	-	-	-	-	97	<1	3	<1	0.08
Antigua and Barbuda	ATG	2017	102	25	97	<1	3	<1	-0.09	-	-	-	-	-	-	-	-	-	-
Argentina	ARG	2017	44 271	92	-	-	-	-	-	-	-	-	-	-	>99	<1	<1	<1	0.06
Armenia	ARM	2017	2 930	63	>99	<1	<1	<1	0.28	>99	<1	<1	<1	0.67	>99	<1	<1	<1	0.07
Aruba	ABW	2017	105	43	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Australia	AUS	2017	24 451	86	>99	<1	<1	<1	0.02	>99	<1	<1	<1	0.06	>99	<1	<1	<1	0.01
Austria	AUT	2017	8 735	58	>99	<1	<1	<1	0.00	>99	<1	<1	<1	0.00	>99	<1	<1	<1	0.00
Azerbaijan	AZE	2017	9 828	55	91	3	5	<1	1.06	82	6	10	2	1.57	>99	<1	<1	<1	0.50
Bahamas	BHS	2017	395	83	99	<1	1	<1	0.04	-	-	-	-	-	-	-	-	-	-
Bahrain	BHR	2017	1 493	89	>99	<1	<1	<1	0.01	-	-	-	-	-	-	-	-	-	-
Bangladesh	BGD	2017	164 670	36	97	2	<1	<1	0.11	97	2	<1	1	0.15	97	1	<1	<1	-0.07
Barbados	BRB	2017	286	31	98	<1	1	<1	0.01	-	-	-	-	-	-	-	-	-	-
Belarus	BLR	2017	9 468	78	96	3	<1	<1	-0.17	98	<1	<1	<1	-0.03	96	4	<1	<1	-0.21
Belgium	BEL	2017	11 429	98	>99	<1	<1	<1	0.00	>99	<1	<1	<1	0.00	>99	<1	<1	<1	0.00
Belize	BLZ	2017	375	46	98	1	<1	<1	0.62	97	1	1	<1	0.76	99	1	<1	<1	0.45
Benin	BEN	2017	11 176	47	66	10	20	4	0.29	58	14	22	6	0.36	76	5	18	1	-0.05
Bermuda	BMU	2017	61	100	>99	<1	<1	<1	-0.01	-	-	-	-	-	>99	<1	<1	<1	-0.01
Bhutan	BTN	2017	808	40	97	3	<1	<1	0.85	97	3	<1	<1	1.08	98	1	<1	<1	0.13
Bolivia (Plurinational State of)	BOL	2017	11 052	69	93	<1	2	5	0.79	78	<1	4	18	1.41	>99	<1	<1	<1	0.26
Bosnia and Herzegovina	BIH	2017	3 507	48	96	4	<1	<1	-0.07	97	3	<1	<1	0.06	95	5	<1	<1	-0.22
Botswana	BWA	2017	2 292	69	90	7	2	1	0.89	76	18	2	4	1.29	97	1	1	<1	0.17
Brazil	BRA	2017	209 288	86	98	<1	2	-	0.27	90	2	8	-	0.92	>99	<1	<1	<1	0.08
British Virgin Islands	VGB	2017	31	47	>99	<1	<1	<1	0.27	-	-	-	-	-	-	-	-	-	-
Brunei Darussalam	BRN	2017	429	77	>99	<1	<1	<1	0.00	-	-	-	-	-	-	-	-	-	-

Introduction | Water | Sanitation | Hygiene

10. The data on drinking water is separated into two sections:
    1) proportion of the population with access to drinking water broken down by service level (safely managed is not included in this section)
    2) proportion of the population using improved drinking water broken down by different dimensions. The data is further disaggregated by rural, and urban populations. See JMP Drinking Water Service Ladder tables for more context.

Table 1 **JMP Drinking Water Service Ladder**	
Service Level	**Definition**
Safely Managed	*Improved* source located on premises, available when needed, and free from microbiological and priority chemical contamination
Basic	*Improved* source within 30 minutes round trip collection time
Limited	*Improved* source over 30 minutes round trip collection time
Unimproved	Unimproved source that does not protect against contamination
Surface Water	No service, open water

Table 2 **JMP Improved Drinking Water Breakdown**

Improved Service Level	Definition
Safely Managed	Water that is located on premises, available when needed, and free from contamination
Accessible on premises	when water is collected from a point located within the dwelling, plot or yard
Available when needed	access to sufficient water when needed, or having water available at least 50% of the time (i.e. at least 12 hours per day or 4 days per week)
Free from contamination	drinking water should meet microbiological and priority chemical water quality standards. The microbiological standard applied is that no E. coli should be detected in a 100 mL sample. Thermotolerant coliform counts are another commonly used faecal indicator, and are considered as an acceptable alternative to E. coli8.
Piped	Piped water into dwelling or Piped water to yard/plot
Non-piped	Can include Public tap, standpipe, Tubewell, borehole, Protected well, Protected spring

11. Let's first explore the data. See if you can determine which countries are without data, and which countries have very low levels of basic water service.
    a. *Hint: Go to the at least basic column under the national level data. Click the filter and select the data with "-" and the lowest number listed (39).*
    b. See what other interesting things you find.

Exercise #2 - Please answer the following questions

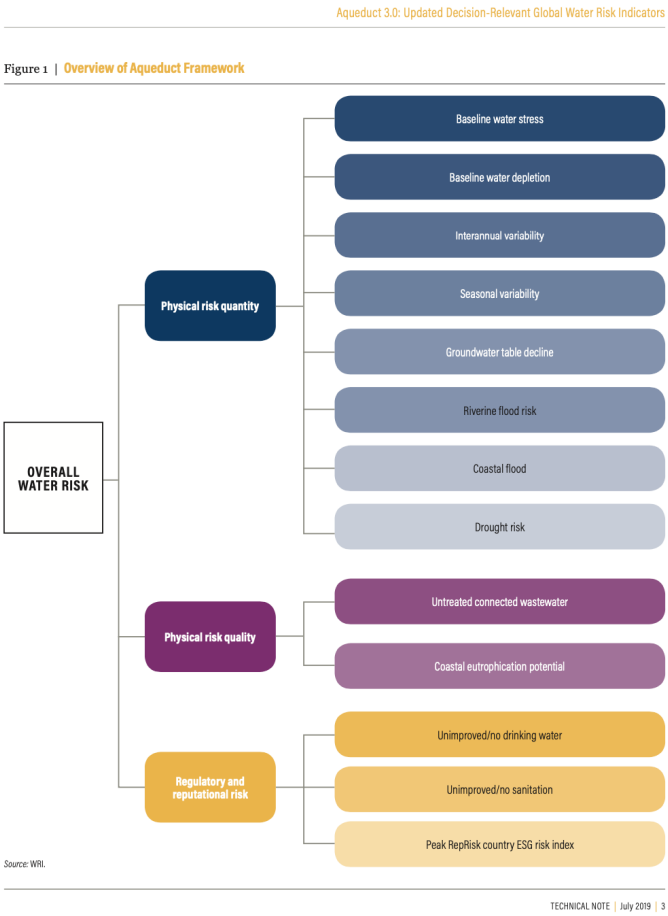
Question 1: What are your initial thoughts on this data set? What questions or concerns do you have after overviewing the data?

Question 2: What country has the lowest proportion of the population with access to at least basic drinking water? Why do you think this country's water delivery is so low? Reflect on the political, geographical, and technical conditions which hinder water delivery.

Question 3: The dash symbol indicates that the JMP does not have data for these countries. What countries have no data on basic drinking water? What about safely managed drinking water? Explain why these numbers differ.

<u>Part 2 Exploring the WRI methodology and data</u>

12. Now that we have evaluated the JMP Data, let's compare it to the WRI data set. The World Resources Institute produces the “Aqueduct 3.0 Decision-relevant Global Water Risk Indicators.” This composite index combines the concepts of physical risk quantity, physical risk quality, and regulatory and reputational risk to create an overall water risk for country and provinces within the country when data is available.
13. Return to the methodology on your desktop (aqueduct-30-Technical Note.pdf)
14. Take a moment to skim through the methodology, paying particular attention to the water risk tree provided on page 3.



Exercise #3 - Please answer the following questions

Question 1: What are the individual indicators that make up the index?

Question 2: Consider the three elements of risk, do you agree with the risk element assigned to the baseline water stress indicator? Do you think any of the other two elements (exposure and vulnerability) correspond to baseline water stress?

Question 3: Page 35 provides an industry specific weighting scheme for the indicators, seasonal variability for the agricultural industry is weighted as 0.5. Do you agree or disagree with this rating? In your answer consider agricultural uses of water and technological capacity of these industries.

Question 4: Name 2-3 ways this dataset differs from the JMP?

15. Let's return to the WRI dataset (aqueduct-30-country-rankings.xlsx)
16. Take a few minutes to familiarize yourself with the data. On the first tab entitled home, WRI provides a link to the WRI data in github, titled "aqueduct30_country_rankings_data_download."[3]
    a. *For those interested this github repository contains an abundance of Aqueduct 3.0 country ranking information for understanding the data.*
17. The dataset has six tabs in total, we will only focus on the "results country" and "results province" tabs.
18. Using the same filter tool used in the JMP data analysis, filter the country data based on its lowest (<10%) risk category (this is the column titled label).
19. Choose one country in particular to explore. Now go to the "results province" tab and observe the provinces in your chosen country.

Exercise #4 - Please answer the following questions

Question 1: Which country did you choose in step 19? Do all the provinces in your selected countries have a low risk?

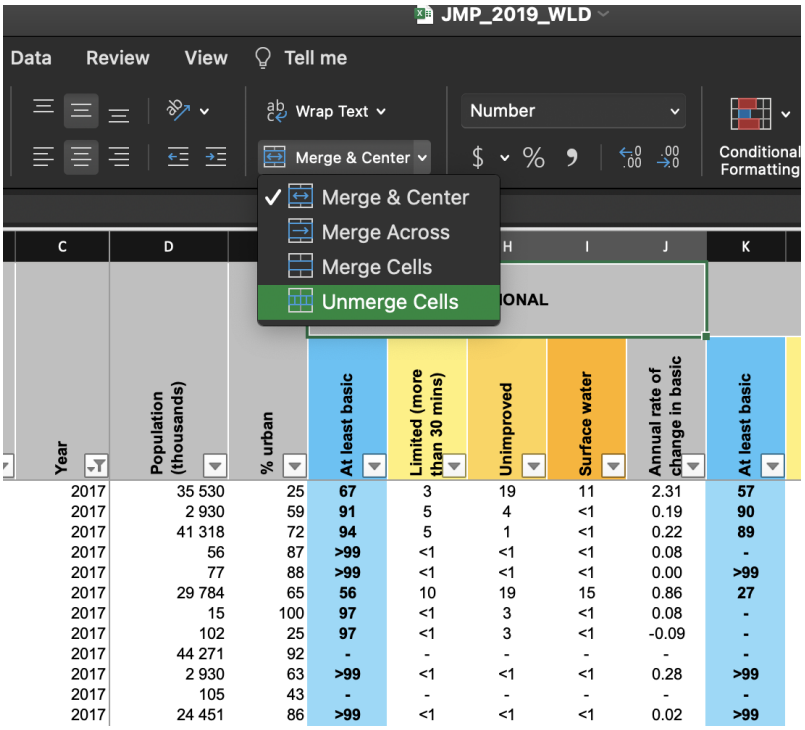
Question 2: Explain in simple terms why risks can vary for each province.

<u>Cleaning the Data</u>

20. Now let's prepare the data for the second part of the module. We need to open a new excel spreadsheet and entitle it "JMP_WRI_Clean." At the bottom of the file, create three sheets labeled - JMP, WRI_Country, and WRI_Province.
21. Head back to the JMP data set and copy the first column in the water tab entitled "Country, Area or Territory". Paste this in the first column of the Clean file in the "JMP" Tab.

[3] If you aren't familiar with github, this link is an online repository for data where you can create a profile and publicly share your data, workflows and results. For more information check out this part of the book.

22. Next from the JMP spreadsheet, we will need to remove the merged cells in the national data column in order to copy the data. Highlight the cell that says “National” and navigate to the toolbar and hover over merge & center and click unmerge cells. Now you can copy the whole column of “At least Basic” and paste it in the second column of the clean spreadsheet. This will be our clean JMP data for the second exercise.



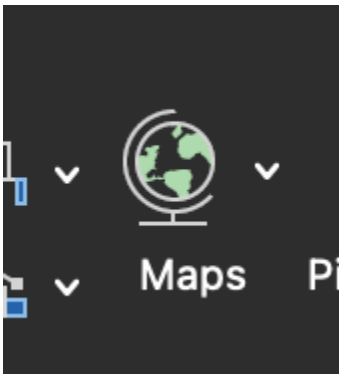
23. Rename the newly pasted JMP column as “%of Population with access to basic drinking water”
24. Now we need to clean the WRI Aqueduct dataset. In the “Aqueduct30-country-rankings” file we will be using the “results_country” and “results_provinces” tabs.
25. First, in the “results_country” tab copy column D (“name_0”) and paste into the clean data spreadsheet under the “WRI_Country” tab. Repeat this copy and paste with column J (“label”).
26. Repeat these steps for the WRI Province tab, copy and pasting columns B, F, and L (column names “name_1”, “name_0” and “label”) into the clean spreadsheet.
27. Download the WRI_JMP_Clean Data from Github, open the file, and compare the data you cleaned to ensure you did so correctly

Exercise #2: Mapping JMP and WRI Data

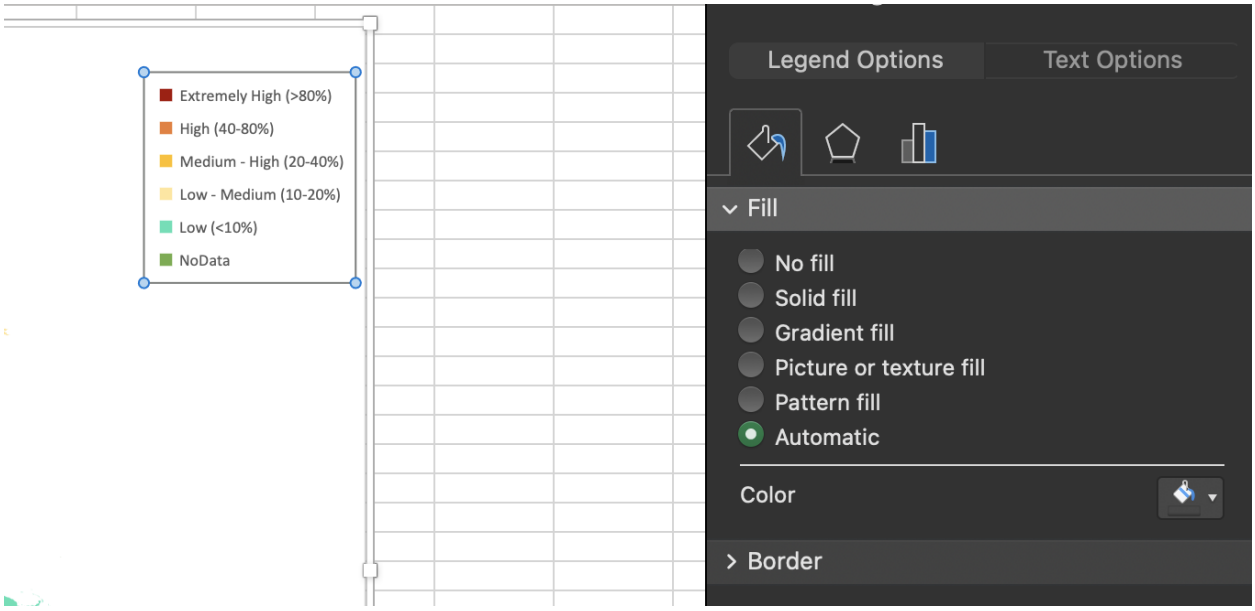
This exercise will take the datasets in Exercise 1 and teach you how to create a map using the data in excel.

<u>Excel Workflow</u>

1. Use the clean data you just created in Exercise 1
2. This file has already been cleaned for you. There are three tabs that we have copied over the data from the WRI and JMP. We have copied:
    a. Country information and associated ranking of water stress for WRI
    b. Country and Province disaggregated data and ranking of water stress from WRI
    c. Country and proportion of population with access to basic drinking water from the JMP
3. We will go tab by tab and create a map for each dataset, showing the data visually.
4. Starting with the WRI Country tab, select the country cell and label columns.
5. Navigate to the top of the window and click insert “Filled Map”.



6. The data will load in map form.
7. Repeat this step for the next two tabs, highlighting country and province data for the WRI Province data set and Country and % of population for the JMP dataset.
8. Change the color coding of each map by double clicking the color on the legend and selecting the fill, and paint can. Below is an example of how we colored the categories for the WRI data.



Exercise #5 - Please answer the following questions:

Question 1: Take a look at each map you have created. Looking at the WRI country map, what areas have the greatest water risk? How does the story change when you look closely at the disaggregated map of water risk by province?

Question 2: Looking at the JMP map, are there any countries or regions that stand out to you for having a high or low proportion of their population with access to basic drinking water?

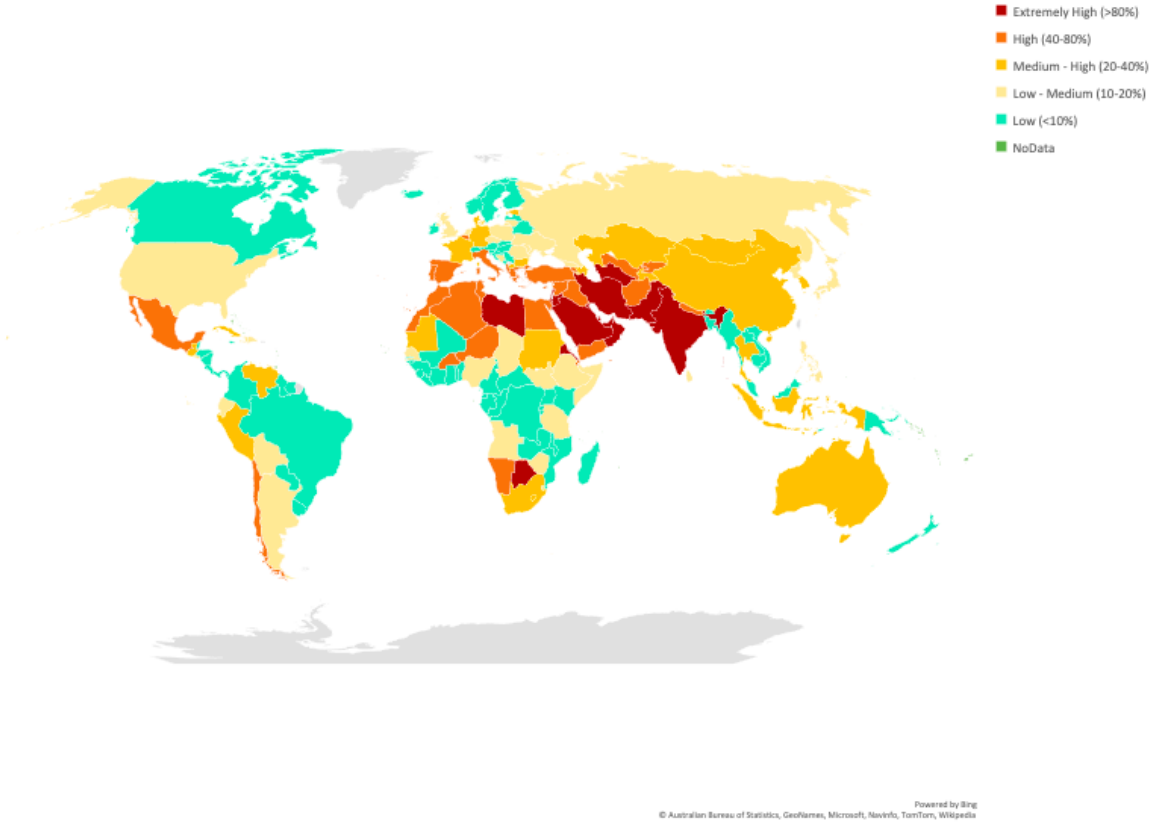
Question 3: Pick a country and compare the WRI Aqueduct water risk ranking to the percentage of people with access to basic drinking water. Are these concepts in conflict or does the data support the conclusion in each?

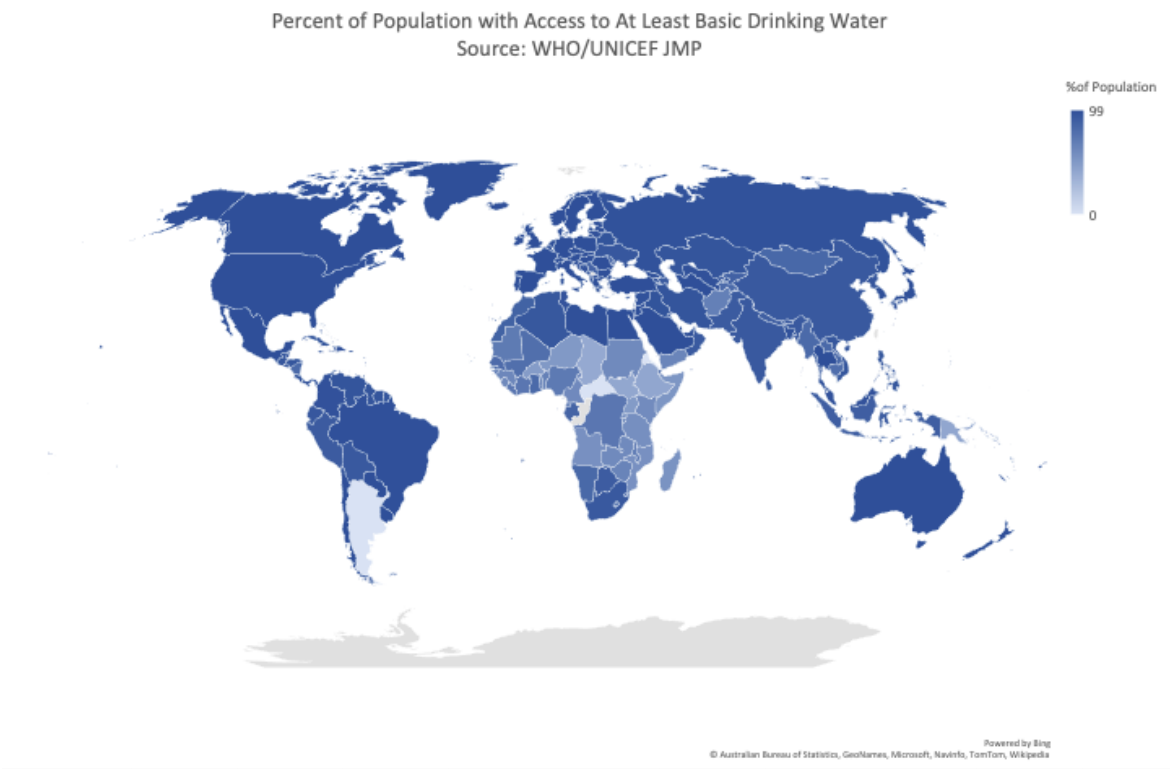
Question 4: Imagine you are tasked with communicating what water access looks like in India to your boss. Which of these maps would you use and why? Under what circumstances would you be more likely to show the WRI map over the JMP?

Question 5: Reflect on the cleaning and visualization process in this data module, at any point did the exercises feel biased? Think about our decision to only map access to basic drinking water from the JMP, do you think that was a good choice?

Map Examples:

Water Stress Ranking by Country
Source: WRI Aqueduct 3.0 Country Rankings





2.3 Useful resources or "For Further Analysis or Discussion"

1. JMP Online Tool
2. WRI Datasets
3. WRI Water Risk Atlas Online Tool