COVID-19 and Democratic Development: Data Cleaning

Team Members

Rachel Kalusniak

Executive Summary of the Final Project

In 2020, COVID shocked populations and stopped life in countries worldwide. The international response to this issue varied. In response to the threat, China implemented strict lockdowns, which forced people to stay in their homes and stopped industrial production to control the virus. On the other hand, citizens in the United States and the United Kingdom exercised their democratic right of free speech to protect stay-at-home orders. The response to the virus also varied by a nation's development level. Countries with robust public health systems could test for the virus and quickly distribute vaccines. Those governments with fewer resources struggled to access personal protective equipment (PPE) and relied on delayed aid from international organizations to distribute vaccines. These trends predict that authoritarian countries with high levels of development can best control future outbreaks.

This project aims to take the first steps to test the hypothesis above, examining the political and health factors that impact a country's control of the pandemic. This information can help aid organizations in predicting where to deploy resources and assist businesses in considering how a disease will affect their operations and workforce. Many auxiliary factors can affect a country's pandemic response. Drawing conclusions to assist in decision-making during the next pandemic is difficult because the data comes from multiple sources that use a variety of conventions. This effort aggregates these sources into a single collection. This project does not aim to do an in-depth statistical analysis or create a prediction model. That is an area for future research. Instead, this takes the first steps to clean and combine the data. This project aims to create a tool with a VBA module that facilitates accessible data transforming and updating. The seamless updates are essential for the COVID data that changes daily. This application allows users to move straight to data analysis, skipping the messing step of data cleaning.

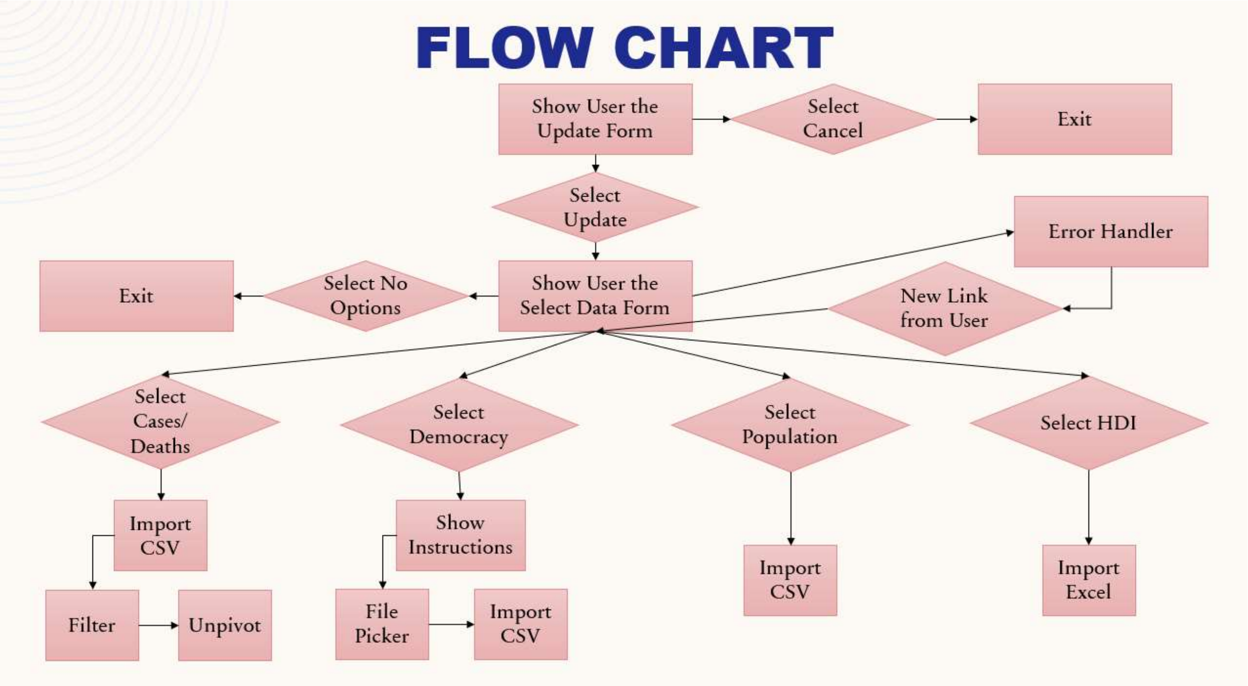
Target Users or Analysis Consumers

This project aims to help future researchers interested in the links between political systems, human development, and pandemic outcomes. This audience will likely be students or academic researchers. It could also benefit international aid programs by assisting them in determining the countries most vulnerable to future health crises. This data will allow the organizations to give extra support to vulnerable countries now to guard against the future. The information is also valuable for businesses as they determine where their workers are more susceptible to disease or their operations are likely to shut down to prevent an outbreak.

Technical Summary

I start examining this project by getting to know the data. I review parts of the democracy data's codebook and other research articles that use these data sources. This understanding helps me grasp the data's idiosyncrasies and shortcomings. For example, the democracy data does not include small countries. The COVID data uses "Diamond Princess" to represent COVID cases for all US citizens, and the HDI data does not include isolated countries like Somalia or North Korea. It is a great geography review. I use the COVID data as primary because it is the largest and the dependent variable. The formulas use arrays and the MATCH function with wildcards to automatically connect as many names as possible in the other data sources. However, this requires some manual entry to assign different terms or remove observations with incomplete data. I aggregate all this work with names and data completeness onto the "CleanNames" spreadsheet, which serves as a codebook for later data analysis. I also created a separate codebook with the data sources and lines to facilitate VBA references later.

Next, I organize the workbook structure to make updating as unobtrusive as possible. I created sheets with raw data with no transformations. This setup means imported data only has to copy and paste into the stagging sheet. Since the COVID data is my primary and very large, I leave it on the import sheet. For the other sources, I rely on the clean names and index match arrays to connect the data from the raw sheets onto one aggregated sheet. This structure includes columns names and years that a future version could update. Finally, the aggregated data is tidy and eliminates extra lookup values.



Once I create a design time structure, I move to VBA coding. The flowchart above shows the logic process when a user launches the update data module. My goal with data importing is simplicity and uniformity. I want modules that could work for multiple data sources. The flow chart shows I accomplish this. Any bottom boxes with the same name reuse the same subroutine. To achieve this flexibility, I create dynamic ranges through formulas like xlDown and xl ToRight and pass variables from the

DataSources sheet to the primary subroutines and the other modules. While there are few user inputs, there is a lot happening behind the scenes as the flow chart shows.

Power queries are the most interesting advanced excel technique I used in this process. This tool allows me to import CSV files from a URL or folder without opening the file. This method increases the processing speed. However, the queries can create destructive issues if the code does not resolve correctly. It is also compelling that a designer must pick a text platform. I use the number 65001 to ensure country names with special characters, like Côte d'Ivoire, import correctly.

The most significant shortcoming of this project is the time it takes to update the data. The unpivot module is the largest source of this delay. I try to do everything for the user with just a few clicks. If I redo the project, I probably will leave the data as it imports and give the user instructions on how to use the power query.

The error handler also is not robust. Its design focuses on links that change, so it could create a death loop if the program has another type of error. This issue was a frustration as I worked on the project. I had to comment out the error handler as I worked through my bugs because it kept skipping there.

There are several areas in this project that I could not complete due to time limitations and their complexity. I wanted to sum the data by country to eliminate the province information. In other programs I work with, this requires a couple of lines of code. I thought fewer observations to unpivot would improve efficiency. However, my efforts with loops and SUMIF struggled and took longer. I read a solution in the textbook to do this using SQL and Access databases. However, that seems like a more extensive idea for a separate project, and I wonder if the already large file could handle it. Additionally, I originally wanted to allow the user to use different years for the input data. For example, they could look at the 2015 population data. The worksheet has prepared formulas for this. However, I do not incorporate it into the VBA code due to the size of the project and the limited usage of historical data for this research question.

Data Needs and Sources

This data relies on six main sources:

- COVID-19 Cases and Deaths (1/2022 – 12/2022)

  This data comes from John Hopkins University’s Coronavirus Resource Center global times series data. This data includes the all-time total for each country. Seeing the daily updates to this data is very compelling as I work through this project. I can tell that the code's regular updates work. The tool imports this data from a URL and copies it into the Excel workbook. I filter the data to include only countries with complete data and run an unpivot. Due to system memory limitations, the data only includes 12/31/2022 to present. This simplification is necessary to make the tool work. However, this tool would make more sense if future iterations of the tool looked at daily cases instead of the current totals.
- Population Data (2022 estimates)

The population data comes from the United Nations' Statistical Division. This CSV file also is a URL to be copied into the workbook. The lookup values filter this data to examine the population data for 2022, but no other transformations are required.

- Democracy Measures Data (2021)

  This data file is from the Varieties of Democracy's (V-Dem) core data. The data set uses various political measures, like the percentage of the population that can vote and the freedom from torture. The project will focus on using the electoral democracy index and liberal democracy index. Electoral democracy is responsive to the citizens through voting, while a liberal democracy protects individual and minority rights against the tyranny of the state. This is the only data the user cannot import directly. Instead, the tool gives users steps and links to download and extract the data. The preparer must delete everything except the first 30 columns to make the sheet small enough to import.
- Human Development Index (HDI) and Gross National Income (GNI) (2021/2022)

  The data originates with the United Nation's (UN) Development Program. This data is a surrogate for a country's public health infrastructure. The assumption is that countries with a higher HDI and GNI have more public health resources. The import Excel feature opens this workbook from the URL, selects the correct tab, and copies the contents to the primary workbook. This data benefits most from the tidying because there are several extra notes and rows. The array formulas also allow sorting this data into very high, high, medium, and low levels of development, a fundamental classification for research discussions.
- Region

  These regions come from the UN Statistical Division's definitions for subregions. This data will allow future researchers to look at performance by region. The information sits on an excel sheet in the workbook. However, this is static data that the user cannot update; This is a safe assumption because countries don't change regions.

Outputs

The code in this module creates a tool that allows users to clean data and import updates easily. It uses a balance of code run time and design time development. The VBA code uses modularization and looping to create flexibility. This tool combines these principles to turn six sheets of messy data in multiple formats and from numerous sources into three tidy data sheets. Future users can easily import this output into statistical or visualization software for further analysis.

Benefits to Target Audience

This tool will save future researchers significant amounts of time. According to an article by the New York Times, data scientists spend 50 to 80 percent of their time collecting and preparing data. This project does that preparation, collecting six data sources into a single workbook and presenting it in a tidy format for easy future analysis. Even if data formats change in the future, the "CleanNames" sheet can be a valuable codebook for users trying to connect country names across sources. The results from this project can give future users that time back, so they can focus on statistical analysis and creating quality visualization.

Challenges

My most significant challenge during this project was the size of the calculations. I became diligent about saving my work because I crashed my Excel or had to force stop the program several times. This project was the first time I saw the warning about my computer being out of memory.

However, the most frustrating part of this project is dealing with queries. This is a potent tool that I use for importing CSV files. It was working fine initially. After working on the error handler, I returned to import another CSV file. I made no changes, but the code was stuck on the ".Refresh BackgroundQuery:=False" line. I spent two hours and half a church choir practice researching the problem. I finally stumbled on the solution when I deleted open queries that failed during error testing and restarted the workbook. Extra open queries do not work with my import CSV method. After this incident, I specifically made one of the first things in my error handler is deleting any open, failed queries to avoid future problems.

As I ran into challenges throughout this project, I relied on great resources to diagnose the problem or discover an alternative solution. I became familiar with great online resources, like Automation Excel and Stacks Exchange. The VBA for modelers textbook also was valuable as I tried to figure out an accessible way to solve my problems.

Personal Learning

This project helped me understand how to pass variables between subroutines. While I used this in homework 4, I struggled with it. With this project, I have an in-depth understanding of how to modularize the code. This ability to pass variables means I only needed one global variable, an array, to move the results from the user form back to the primary module.

I also gained advanced looping skills throughout this project. I started writing everything out but then recognized the applicability of a loop and the correct format. The code includes all three types of loops: For/Next, Do While, and For/Each. I learned that VBA code could step backward in a loop. This technique is essential when deleting things. Otherwise, the shifting numbers can lead to the program deleting all data. I made this mistake once. However, the importing feature means any errors were easy to start over from with a fresh data pull.

Overall, this project taught me persistence and problem-solving skills. There are a lot of resources about VBA coding, but the designer is ultimately responsible for making it works or creating error. Sometimes I wanted to bang my head against the table, but I learned how to read error reports and diagnose problems. Ultimately, I got the program to work.

Closing Thoughts

An outbreak cannot catch the world unaware again as it was with COVID-19. Governments and international organizations must work together to reinforce vulnerable areas, but this is only possible when data is accessible to decision-makers. This project aims to create a comprehensive data set and user dashboard that examines COVID outcomes related to political and health factors. This effort will allow for future robust research testing the hypothesis that highly developed autocracies are the most protected against pandemic threats. This project does not judge the morality of pandemic controls; it just looks at the data.

This project greatly improved my ability to work in VBA. However, this project also showed the shortcomings of Excel. Other programs could have handled this load better. I am proud of what I accomplished with this tool.