Data Analytics Bootcamp Project 2 - World’s Statistical Capacity

Members: Claude David, Zach Wenz, Alexander Mould, Bill Malley

Abstract

The objective of this project was to extract data from the World Bank Group and the World Health Organization and compile them into a central database. Using this database, one can conveniently compare every developing country's statistical capacity to their nation’s preparedness to unforeseen emergencies.

Introduction

Good data are key to reducing poverty, improving health, political stability, and strengthening economies worldwide, including in the United States. There are many countries that lack in collecting, analyzing, disseminating high-quality data about their population and economy. The capability of a country to record, analyze, and store data is referred to it’s “statistical capacity”. It is important for every nation to be able to collect data because quality statistics are essential for all stages of evidence-based decision making, including:

* Monitoring social and economic indicators, including progress toward the Millennium Development Goals (MDGs) and Sustainable Development Goal (SDG)
* Allocating political representation and government resources
* Guiding private sector investment
* Informing the international donor community for program design and policy formulation

In 2010 an earthquake of magnitude 7.0 hit the capital of Haiti and killed nearly 230,000 people died in a matter of seconds. In the aftermath of the quake, efforts by citizens and international aid organizations to provide medical assistance, food, and water to survivors were hampered by the failure of the electric power system (which already was unreliable), loss of communication lines, and roads blocked with debris. A week after the event, little aid had reached beyond Port-au-Prince; after another week, supplies were being distributed only sporadically to other urban areas. Due to Haiti’s low statistical capacity of 53.3, there was not enough data about the country’s population thus no one knew where to efficiently distribute all the aid. This is a perfect example of how the quality of every country’s statistical capacity could be useful. This project strives to infer that the more that a country is able to collect, analyze, and store data the more prepared they are to unforeseen circumstances.

This project was divided into three main components: Extract, Transform and Load. First, data was extracted from two different websites using the BeautifulSoup and Pandas library, transformed into the required format using the Pandas library and then loaded into PostgreSQL. The analysis is done also using the Pandas library. Finally, plotting is done using the Matplotlib library.

The questions that this project proposes to answer are:

1. How prepared are developing countries to unforeseen circumstances based on statistical capacity?
2. How can we use statistical capacity to help countries in need?
3. How can statistical capacity be used as a basis for a country's disaster/outbreak.

Resources

The websites that are used in this project are:

1. <https://data.worldbank.org/>
2. <http://apps.who.int/gho/data/view.main.IHRCTRY05v>

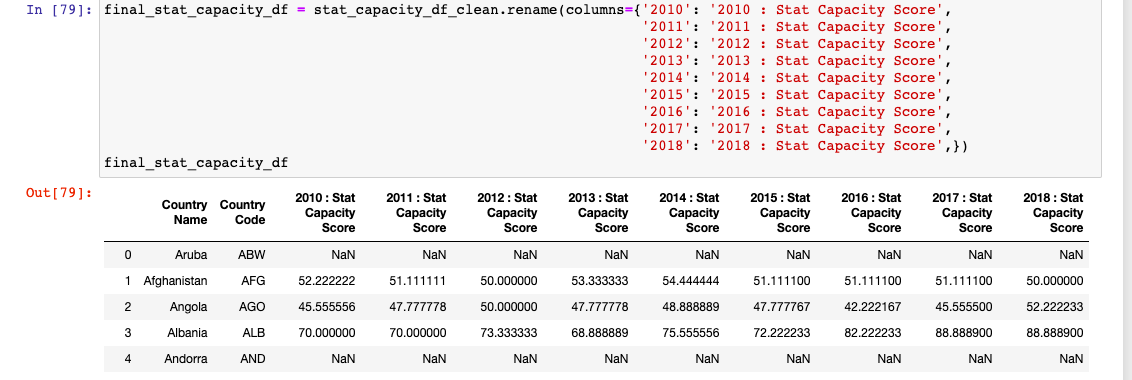
Technologies used are:

1) Jupyter Notebook (Python 3 and Python Libraries like Beautiful Soup, Pandas, and Matplotlib)

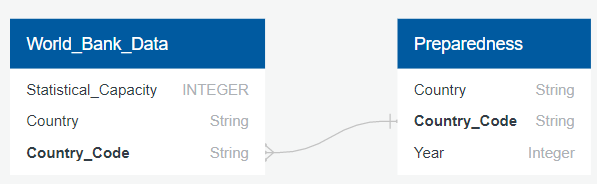
2) PostgreSQL

Methods

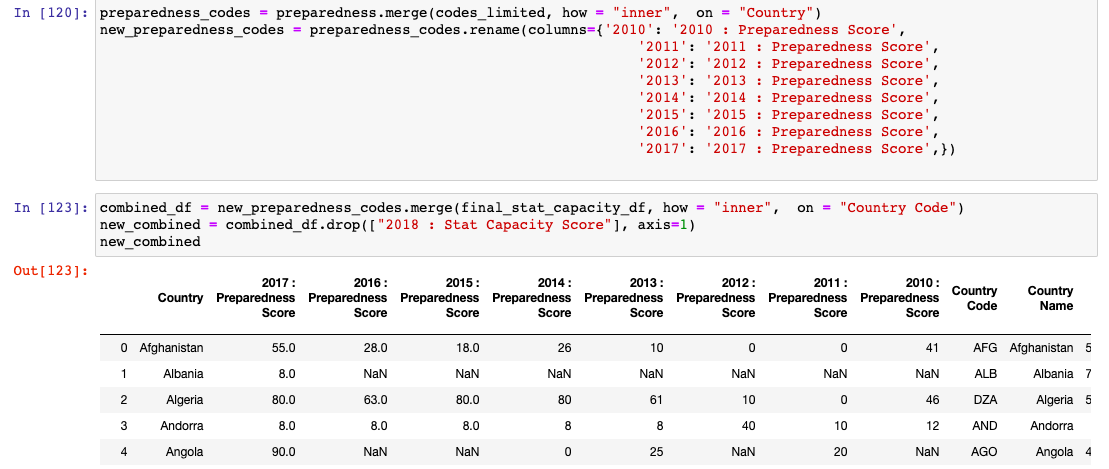
**Extract:** We extracted data from the World Bank Group and gathered the statistical capacity values for all the countries listed. We then read the CSV into Pandas then dropped all columns that have the years less than 2010 and renamed the columns.



We then created an Entity Relational Diagram to have a better flow for the remaining of the project

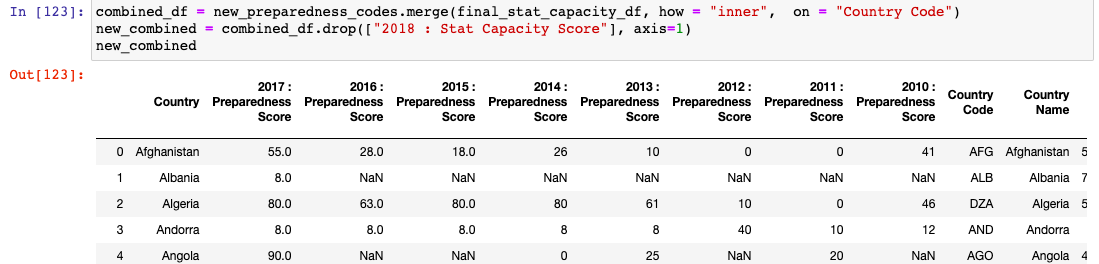


We then extracted data from the World Health Organization and the world Bank group and gathered each country's preparedness and country code from two different CSV files. We pulled the columns that we are interested in from the country code csv then did an inner join with the preparedness table.



We then added a column for the Wikipedia description of each country’s census collection process and history, if available, by scraping text from: <https://en.wikipedia.org/wiki/Population_and_housing_censuses_by_country#cite_note-4>.

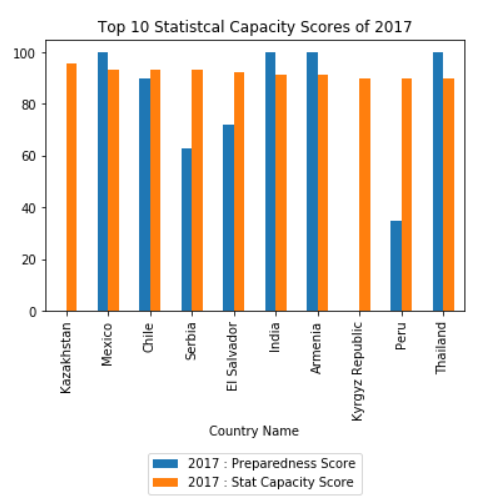
**Transform:** We then transformed our data by doing some cleaning, renaming, and merging different data frames then saved our new CSV. This was our final data frame:



**Load:** We loaded our relational database into PostgreSQL using Pandas

Opportunities for Analysis

The World Health Organization or other international nonprofits can benefit by using this data to determine the potential impact of aid dollars that are specifically directed towards raising a country’s statistical capacity, potentially through census collection efforts. By examining the relationship between a country’s ability to collect and disseminate data and their disaster preparedness score, it can be determined which countries would benefit most from aid dollars directed towards increasing their census collection efforts.



Comparison of top 10 statistical capacity score to Preparedness score

**Reproducible ETL steps**

1. Choose data to extract.
2. Read in the data and place it into a dataframe.
3. Group the Data Frame by most relevant data.
4. If there is no obvious choice for a data table merge. Create a normalized category (bins) for joins, merges and comparisons.
5. Merge the normalized category into the data tables.
6. Repeat steps 1-5 for every imported data set.
7. Clean the data: a. Remove duplicate data. b. Drop incomplete data records. (for this project any 1st world countries or countries with preparedness ≥ 90 were dropped). c. Standardize data types. (Language, date/time, encoding, etc.) d. Remove any files with a count beneath your threshold of interest to lessen data static.
8. Create a Entity Relationship Diagram (ERD)
9. Create new data frame(s) with the desired columns in the desired order.
10. Compare new data frame(s) to the original dataframe to compare category lists.
11. Merge tables and set normalized categories as primary keys.
12. Remove any missing values once more to organize data into logical order.
13. Save the final dataframe to a csv (or prefered file option).
14. Load the file into a relational SQL database for future business or analytical use.
15. Visualize data and present as needed.