Global Temperatures & Health

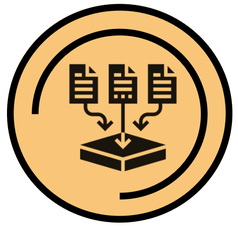
AN ANALYSIS OF CHANGING CLIMATES AND GLOBAL HEALTH

Colin Phillips
Adnan Desai

Introduction

EXPLORATORY QUESTIONS

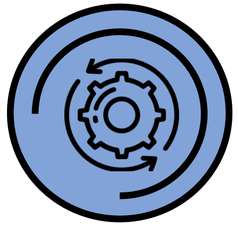
Is there a correlation between a rise in global temperatures and declining global health? What speciic health metrics and demographic statistics are correlated with rising temperatures? What regions have been impacted the most?



DATASETS

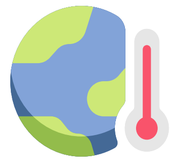
The datasets used were obtained from Kaggle:

- Global Surface Temperatures (GST), 1743 to 2015
- Health Nutrition and Population Statistics (HNPS), 1960 to 2015



PROCESSING

The above datasets were loaded into BigQuery, with five initial tables from GST and a single table from HNPS. A combination of SQL and Apache Beam transformations were used to generate a total of 41 intermediate and modeled tables. Workflows were automated using Apache Airflow.



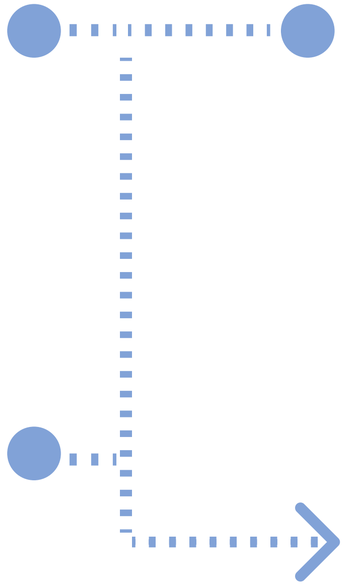
Global Surface Temperatures

INITIAL TABLES

The following tables were stored in kaggle_staging:
-> Global_Temperatures
-> Global_Land_Temperatures_by_State
-> Global_Land_Temperatures_by_Country
-> Global_Land_Temperatures_by_Major_City
-> Global_Land_Temperatures_by_City

TRANSFORMATIONS

The initial tables were created using SQL to pull from the appropriate files uploaded in a GCP bucket; casting performed to ensure correct types. The major city and city tables were combined into a singular City table. Once the tables were created, dates were filtered via Beam pipelines to ensure only dates in the range of the Date table were included to ensure no foreign key violations. Additionally, duplicate records were removed from the City table, using beam.

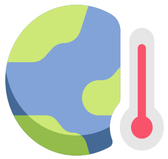


BEAM FUNCTIONS

Functions used within Beam pipelines:
-> **FilterDateFn()**: ParDo function that filters out any dates before 1755
-> **GroupByKey():** Built-in transform that groups PCollection based on keys
-> **DedupCityRecordsFn():** ParDo function that returns first element in grouped PCollection to remove duplicates

MODELED TABLES

Final tables were stored in kaggle_modeled:
-> Date
-> State_Beam_DF
-> Country_Beam_DF
-> City_Beam_DF



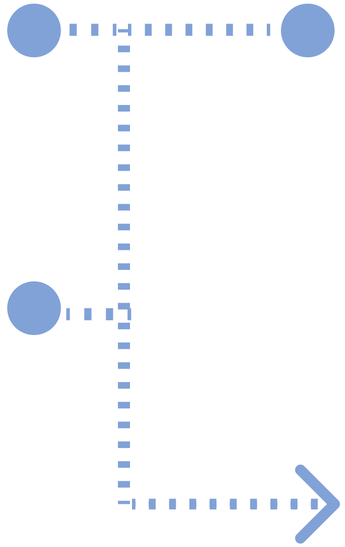
Health & Population Statistics

INITIAL TABLES

The following table was stored in kaggle2_staging:
-> Health_Nutrition_Population_Statistics

TRANSFORMATIONS

The initial table were created from pulling certain relevant statistics through their metricCodes via SQL and grouping them in four tables: Population_Statistics, Urban_Growth_Statistics, Life_Statistics, and Health_Statistics. For modeling, each of the tables were transposed via a Beam pipeline such that each of its date columns were now a record with a singular date attribute. This allows each record to represent a single day for cross-dataset analysis.

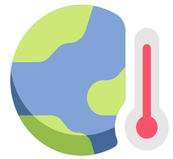


BEAM FUNCTIONS

Functions used within Beam pipelines:
-> **TransposeDateFn()**: ParDo function that takes each date column and creates a record for each while redefining the schema
-> **generate_elements():** A FlatMap transform that flattens the transposed dates list to separate the records

MODELED TABLES

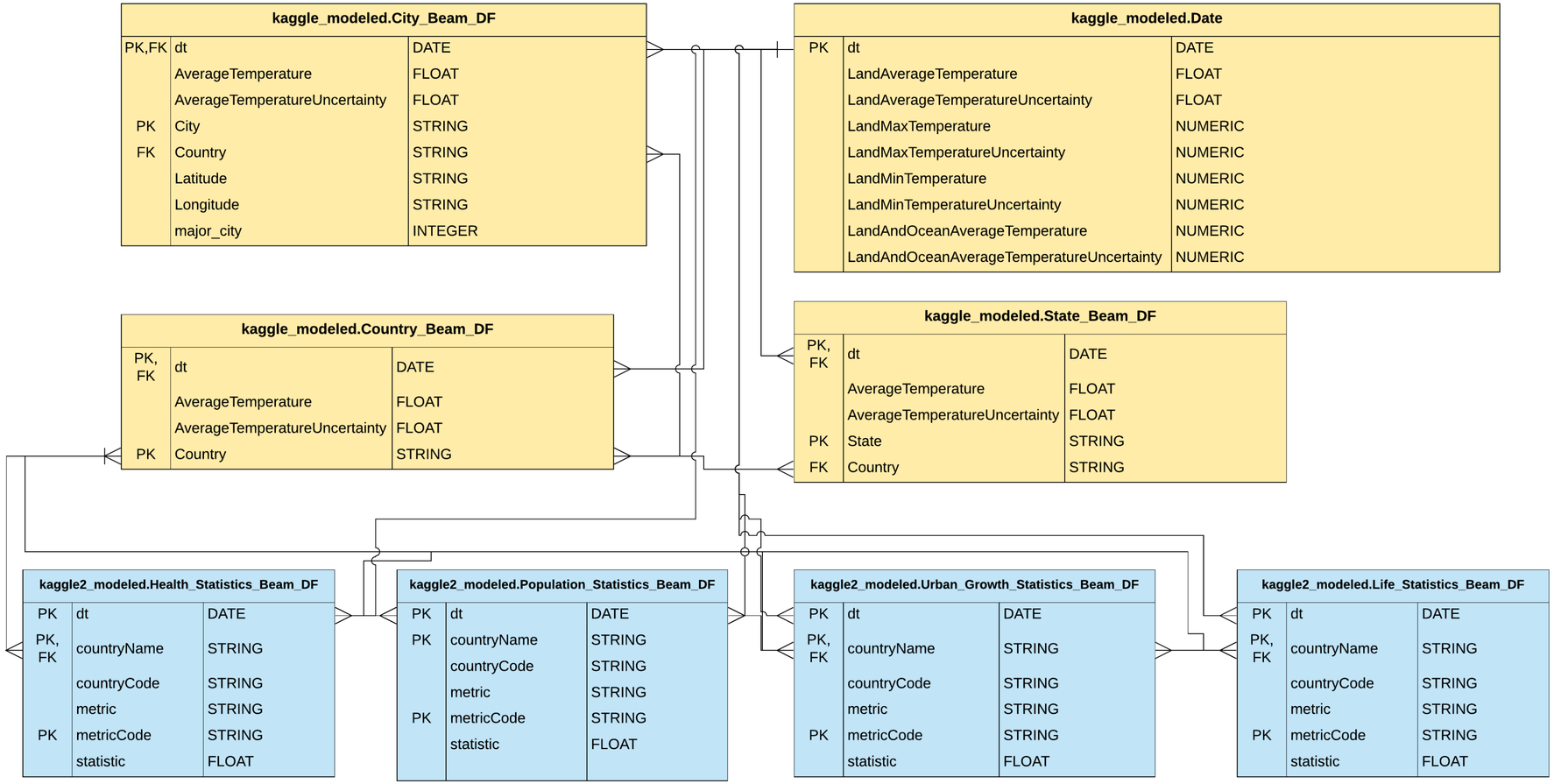
Final tables were stored in kaggle2_modeled:
-> Health_Statistics_Beam_DF
-> Life_Statistics_Beam_DF
-> Population_Statistics_Beam_DF
-> Urban_Growth_Statistics_Beam_DF



Modeled ERD

DATABASE DESIGN

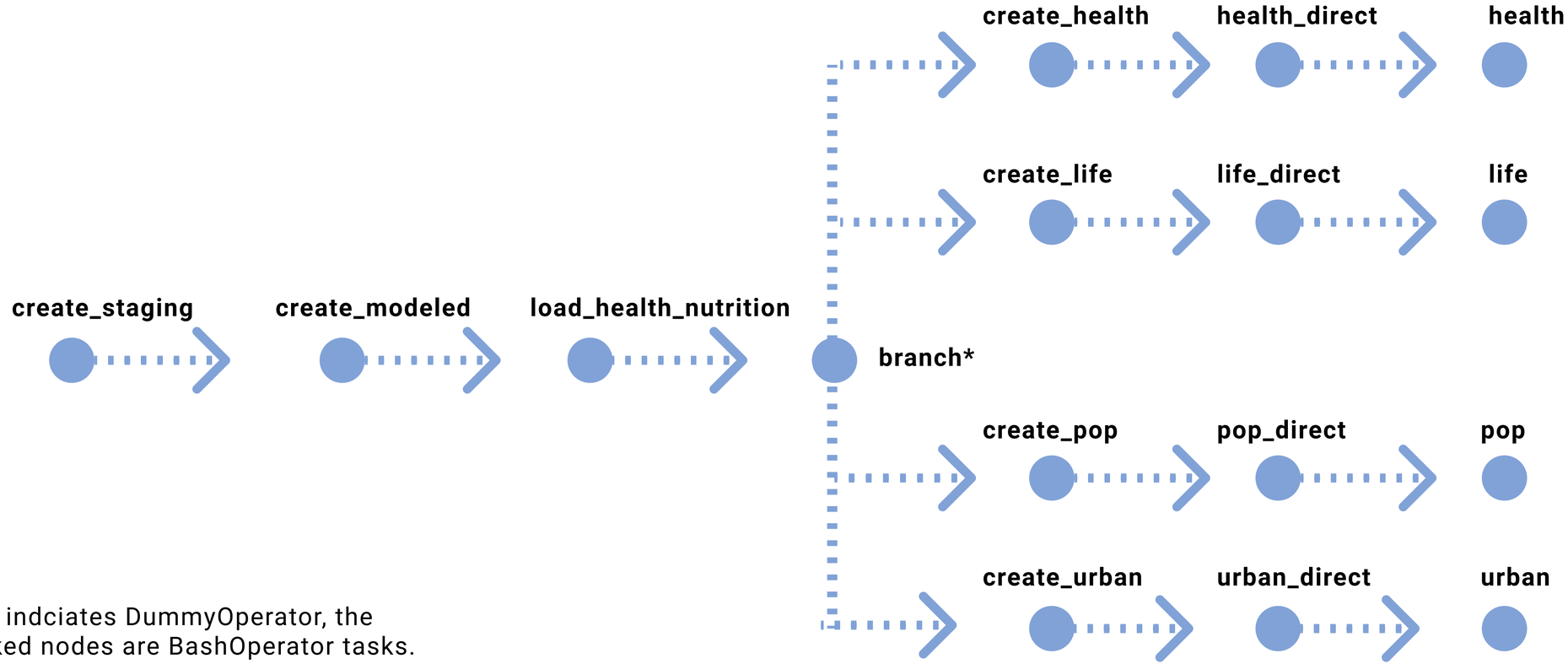
Following the creation of the eight modeled tables, an Entitiy Relationship Diagram (ERD) was finalized to model the relationships each table has with the other, including across original datasets. In most instances, all tables except Date contain many-to-many relationships with the others. Referential violations were checked and corrected with the Beam pipelines, with each table having a primary key and possibly multiple foreign keys.



Workflow Automation

WORKFLOW DAG

A directed acyclic graph (DAG) was created to automate the secondary dataset workflow using Apache Airflow. Branches were used to allow for parallel execution of independent tasks. The produced tables were stored in kaggle2_workflow_staging and kaggle2_workflow_modeled in BigQuery.



Note: * indciates DummyOperator, the unmarked nodes are BashOperator tasks.

Cross-Dataset Analysis

JOINING WITH COUNTRY TABLE

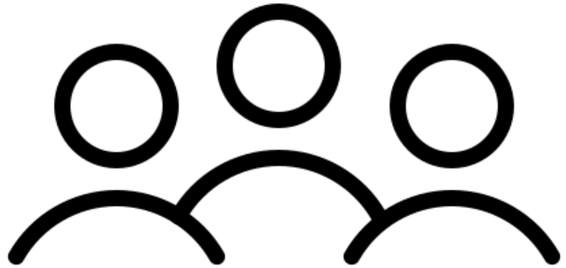
In order to effectively join the datasets, the transposed date columns were used from the tables in the secondary dataset. With this feature, three cross-dataset queries were implemented by joining the Country entity table with each of the secondary dataset tables.

HEALTH STATISTICS

This cross-dataset query visualized the relationship between country temperature and male survival to 65 percentage over time.

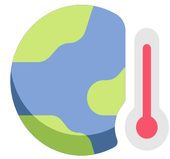
LIFE STATISTICS

This visualization displayed the relationship between country temperature and crude death rate.

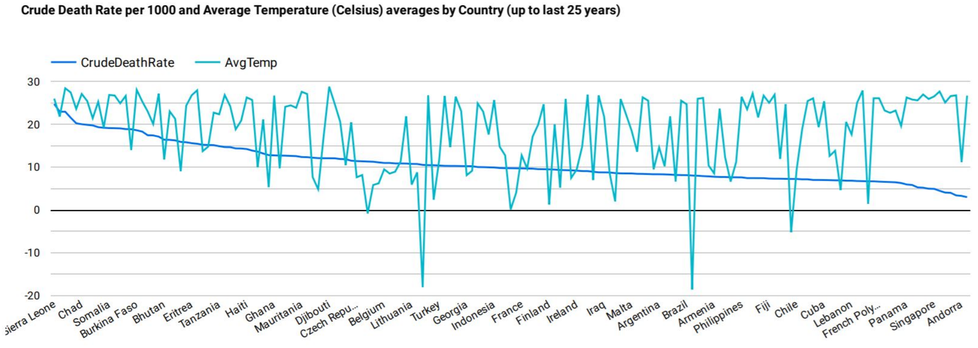


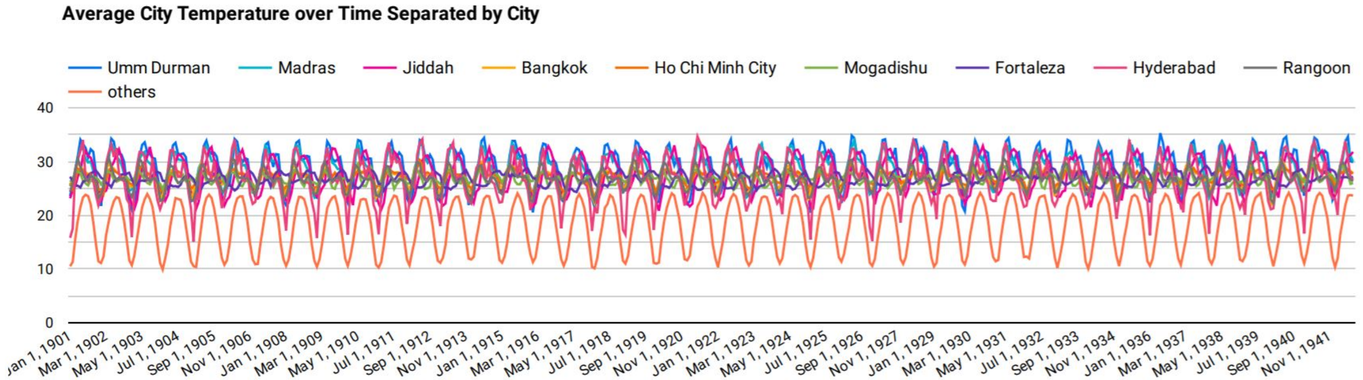
POPULATION STATISTICS

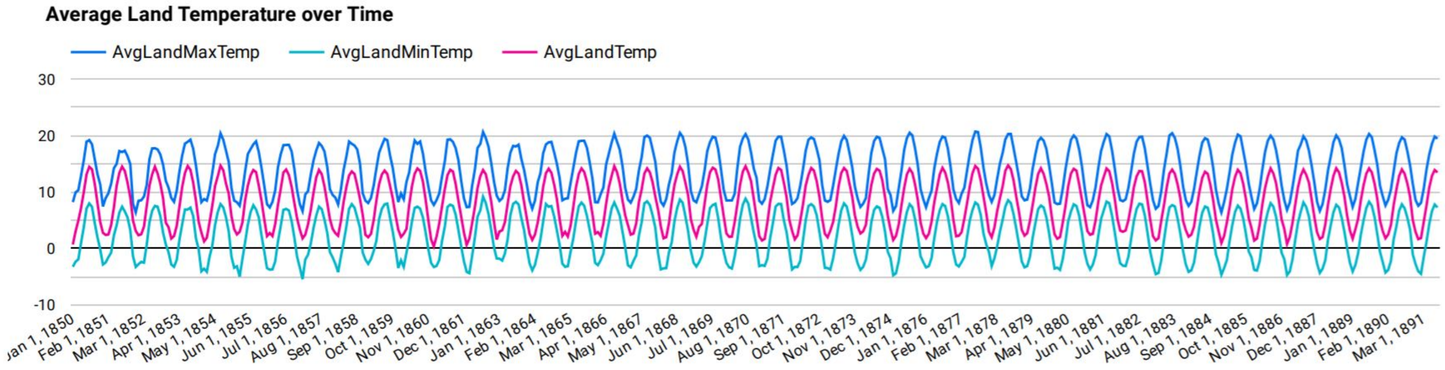
This query visualized the relaitonship between country temperature and the total population of India. This will be expanded for other densely populated countries in the future.

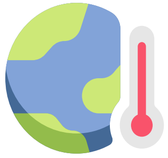


Visualizations

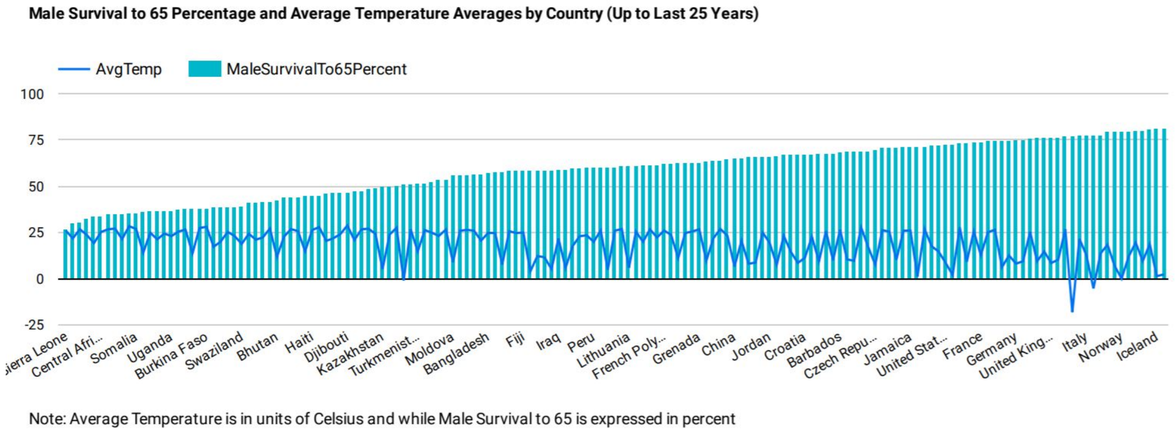




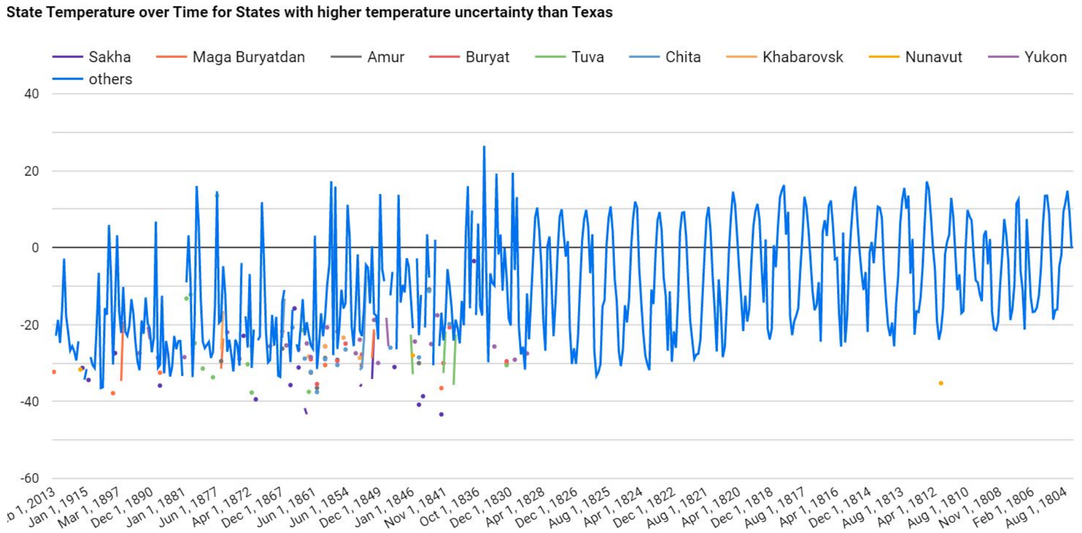


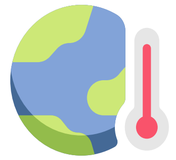


Visualizations



Note: Average Temperature is in units of Celsius and while Male Survival to 65 is expressed in percent





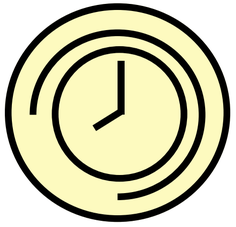
Model Improvements

DIVERSIFY METRICS

Of the over 160 metrics included within the original dataset for Health Nutrition and Population Statistics, the current secondary dataset only contains information for 24 of them. Inclusion of additional metrics would allow for discovering better correlated metrics and the possibility of classifying these metrics beyond the four current entities. These may involve adding additional tables to the database.

EXPAND TIMEFRAME

The primary dataset spans from 1743 to 2015, though not all countries, for example, have entries dating that early. To improve insights, queries can be performed to determine which countries have the earliest available temperature data. Health metrics of interest could then be identified and then scraped to expand the secondary dataset timeline to the 18th century. It is certainly limiting to only be able to perform cross-dataset analysis for 65-year period.



SCOPE EXPANSION

Originally, the goal was to have the secondary dataset be focused on global particulate matter. Unfortunately, there was limited data available that typically started in 1990. Further research could yield a verified source with a comparable time period as the primary dataset. Combined with global temperature information and health metrics, this could be an interesting addition that would expand the possible insights that could be obtained from the database.