Rohit Prakash, Tanya Thaker

Design Decisions

There were a couple of changes we made to our schema based on the feedback we received on our phase 1 submission.

The first comment suggested that we split our table containing the GDP and GDP per capita (PC) for countries in a given year into two separate tables, one for GDP and one for the per capita. The initial table had an attribute representing which country and year the row was about, an attribute showing whether the row was for the GDP or the PC, and finally the actual money amount. The reasoning for this was to try and combine tables where we could to reduce redundancy although we weren't sold on the idea. Now with the feedback we feel more comfortable having two tables, one for the GDP and one for the GDP PC.

There was a similar piece of feedback regarding our TB_by_country table, with the TA suggesting not only to split this table up as well, but to change our primary key. Our decision was to split the table into two smaller tables, each with its own focus. We put all this information in the TB_by_country table, but it resulted in redundancies in certain values, so to fix this we made one table focusing on the rates of TB overall, and another focusing specifically on how different demographics were affected by TB. The redundancy was caused by, amongst other things, having the overall rates be the same for multiple rows with the same country and year because the demographic split was different.

There were some additional design decisions we made once we more closely analyzed the data and figured out exactly what we could represent and how the data should be structured to represent it.

We created three variable types that are used by our tables to only allow for values in those types to be inserted. This was done to constrain the data and make sure only expected and cleaned data was accepted.

Most of our tables have a foreign key reference to the TB_Rates table primary key, as the focus of this schema is to look at different aspects of tuberculosis infections. The primary key for the TB_Rates table is country, year (identifying a particular time and place) and most of our other tables have the same attributes so we can compare different aspects based on countries or years. Our data was therefore limited to information about countries that had TB data collected on them in the years found in our source dataset, so even though we had larger datasets to base our schema around, we parsed it down to match the TB information we had.

The only table without a foreign key to TB_Rates was the diseaseCausedDeaths table, which is because the information isn't limited by country, only by year. As the year attribute is not unique nor the primary key for TB_rates, we chose not to reference it.

We also knew that there would be missing data, and allowed for that in our design. For example, not every country allocates funds to every healthcare sector, so rather than

having each sector as an attribute (potentially needing nulls) we have an attribute describing the sector, and another to show the amount.

Cleaning Process:

From all the datasets we investigated we only found common year variables for 2019 and 2020. Thus we cleaned every table for years 2019 and 2020. Furthermore we ensured that we had information only for countries for which we knew the TB fatality and incidence rate. Observations with null values were removed.

- GDP Table
  The dataset from WorldBank had values of GDP for years from 1960 to 2020 in the corresponding year columns. So we created a column 'country' with all countries in the dataset. Each country had two observations for GDP, one for 2019 and the other for 2020. Thus the column 'country' had each country name repeated twice, once for 2019 and the other time for 2020. Then, we merged the columns 2019 and 2020 from the original dataset and added the column to the cleaned dataset. Lastly, we removed all fields with NULL values.

- GDP Per Capita Table
  This table was cleaned in the same way as the GDP Table above.

- TB_Rate
  From the dataset we chose 4 variables: 'country', 'year', 'e_mort_num', 'e_inc_num' which corresponds to country , year, total mortality, total TB rate. We filtered for only 2019 and 2020 and ensured the absence of null value observations.

- TB_Demographics
  From the dataset we chose the variables we wanted to investigate i.e. country, year, risk_factor, age_group. We filtered for countries that exist in the TB_Rate table. The dataset contained some ambiguous age_groups which we removed while we converted the remaining to fit our domain. The conversion was as follows:

Initial Dataset	Cleaned Dataset
0-14	removed
0-4	0-4
15-24	15-23
15plus, 18plus	15plus
25-34, 35-44, 45-54, 55-64, 65plus	15plus
all	all

- DiseaseCausedDeaths
  The table was easy to clean since we just had to combine data from different sheets. No additional cleaning and filtering was required.

- HealthCareFunding
  There were two datasets for this table one for 2019 and the other for 2020. Each of them had columns for each of the healthcare sectors representing the amount received in the corresponding sector. There were 6 healthcare sectors we were considering : “Drug Susceptible TB”, “Drug Resistant TB”, “Research”, “Patient Support”, “Budget Line” and “Lab”. We created a column ‘country’ that contained country names, “year” with the value 2019, “health sector” representing the sector and “amount” representing the funding received. Each ‘country’ was repeated 6 times in order to store the possible health sector for each country in 2019. Then, we merged columns from the original dataset for funding and appended it to the cleaned dataset. Finally, we removed the NULL values.
  The same process was repeated for the dataset representing values for 2020. Then, we binded the two cleaned datasets by rows.