Physical Design and Data Staging

CSI 4142 - Fundamentals of Data Science

Winter 2022

School of Electrical Engineering and Computer Science

University of Ottawa

Course Coordinator: Dr. Herna L Viktor

Teaching Assistants: Anuhbav Chharbra, Nicolas Fleece, Paul Mvula

Group 25:

Lilian Ly, 8262186

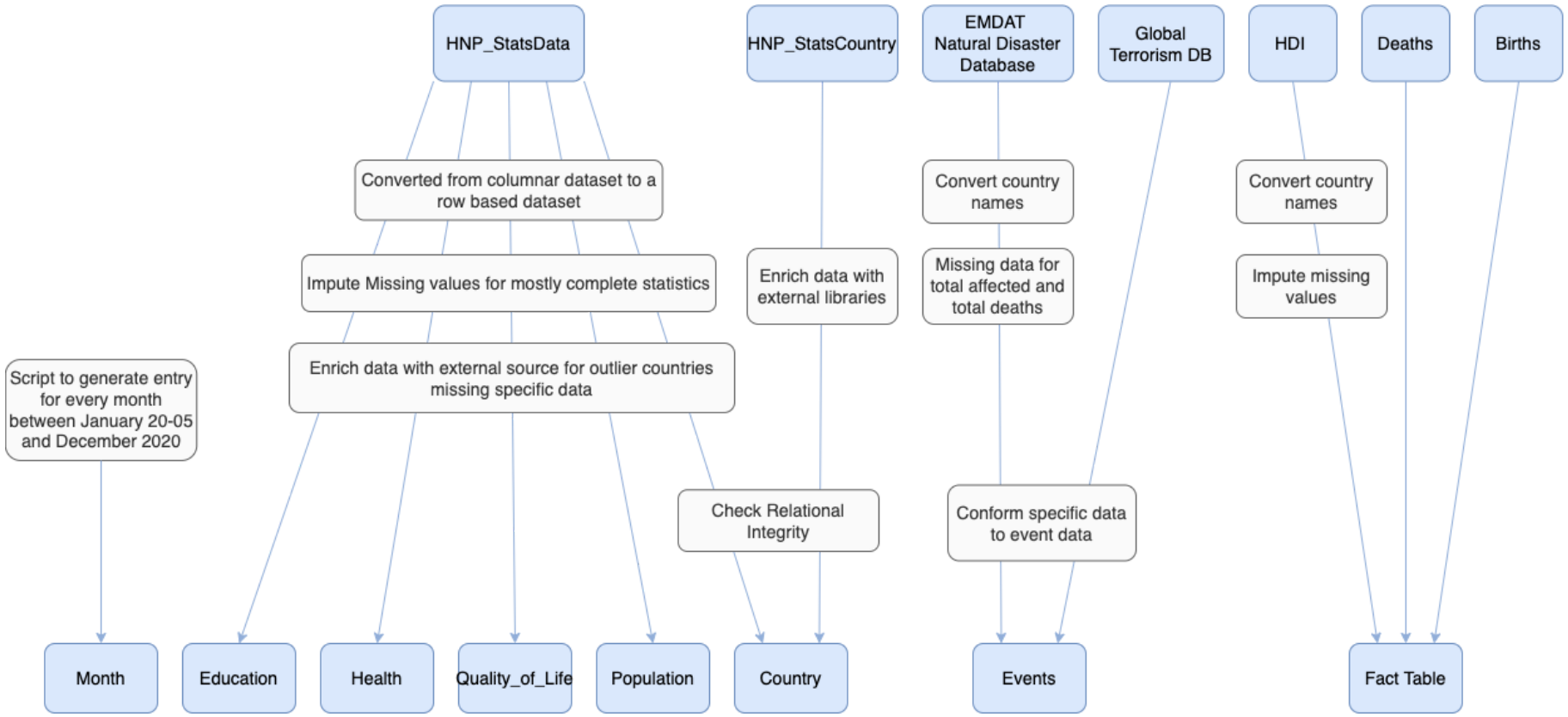
Jonathan Brar, 8209351

Logan Rose, 300059034

Submission Date: March 19, 2022

Table of Contents

High-level Data Staging Plan



Data Quality Issues

Education Dimension, Health Dimension, and Quality of Life Dimension

We handled missing data by manually filling in the information in the *HNP_statsdata.csv* file from external sources found online. In some cases where data could not be found, we took the mean values of previous years and used that for data from 2005 - 2020. Missing data was detected when we ran our Pandas scripts to retrieve our attributes from the databases. The attributes that had empty spaces in the resulting csv files were missing data. We attempted to find data that was in between 2005-2020 and manually inserted it into the larger *HNP_statsdata.csv* file. Our pandas scripts then pulled the data from this csv file and filled out the missing areas. The following list is the different attributes that were missing data for the specific country.

- Canada
    - Literacy Rate
        - Canada Literacy Rate 1990-2022 | MacroTrends
    - Poverty headcount
        - Canada - Poverty Headcount Ratio At $1.25 A Day (PPP) (% Of Population) - 2022 Data 2023 Forecast 1971-2017 Historical (tradingeconomics.com)
    - Total Primary Completion Rate
        - Canada - Primary Completion Rate, Total, Based On Completers - 2022 Data 2023 Forecast 2012-2016 Historical (tradingeconomics.com)
    - Female Primary Completion Rate
        - Canada - School Enrollment, Primary, Female (% Gross) - 2022 Data 2023 Forecast 1971-2019 Historical (tradingeconomics.com)
    - Male Primary Completion Rate
        - Canada - School Enrollment, Primary, Male (% Gross) - 2022 Data 2023 Forecast 1971-2019 Historical (tradingeconomics.com)
- USA
    - Poverty headcount
        - United States - Poverty headcount ratio (indexmundi.com)
- Brazil
    - Total Primary Completion Rate
        - Mean value of previous years used
    - Poverty headcount
        - Brazil - • Brazil - poverty headcount ratio 2005-2015 | Statista

Events Dimension (Disaster, Terrorism)

Two external sources were used to create the Events dimension; quality issues were handled differently for each source:

- **Disaster events** (*emdat_public_2022_02_31.csv)*:
  The *emdat_public_2022_02_31.csv* contained missing values in the total affected and total deaths columns. We did some research and concluded that the empty spaces meant that there was either 0 total affected or 0 total deaths. So if there was a value greater than 0 in total affected and no value in the total deaths, then the total deaths value was set to 0. If there is a value greater than 0 in total deaths and no value in total affected, then the total affected value is equal to the total deaths value. If there was no value in either column, then both were set to 0.

  The second issue with this data set is that some disasters were missing information in the date columns. If this data was missing in a row, that entire row of data was removed from the seed file. We had plenty of disaster events and felt it was fine to remove a few rows if they were missing date information.

- **Terrorist events** (*globalterrorismdb_0221dist_filtered.csv)*:
  The large data source contained missing and null values under the attack type (*attacktype1_txt*), property damage (*propextent_txt*), number of wounded (*nwound)* and number of deaths (*nkill*). As only 10 events are required per country, and there were a total of 44 thousand terrorist events found for the 9 countries, such events that had any missing or null values were ignored and dropped from the data set. The data was further narrowed down by choosing only events that had an impact on the population number by taking events that had at least 1 death or wounded person.

Integrating the data to form one single event dimension from the two data sources stated above was done by first renaming all columns to match from both sources. The columns were then consolidated into being: Key, Seed_id, Country, Start_year, Start_month, Start_day, End-year, End_month, End_day, Event_type, Total_affected, Total_deaths. By doing so, both sources were merged through concatenation into one single *Event_seed.csv*.

Human Development Index (HDI) Measure

One of the measures is the Human Development Index (HDI) taken from the data source *HDI.csv*. Since the data time frame is taken between 2005-2020 (inclusively), there was missing data for all of the year 2020. The missing data was handled by taking the last most updated HDI for each country which was provided in 2019 to be the same for 2020.

Team Work Plan

Deliverable Checklist	Responsible Team members	Expected completion date	Actual completion date	Estimated Time to complete	Actual time to complete
Create db instance	Logan	March 5th	March 5th	1 hour	1 hour
Create Country dimension	Logan	March 5th	March 5th	1 hour	1 hour
Create Month dimension	Logan	March 5th	March 5th	1 hour	1 hour
Create Education dimension	Logan, Jonathan	March 5th	March 5th	1 hour	1 hour
Create Heath dimension	Logan, Jonathan	March 5th	March 5th	1 hour	1 hour
Create Quality of Life dimension	Logan, Jonathan	March 5th	March 5th	1 hour	1 hour
Create Population dimension	Logan	March 5th	March 5th	1 hour	1 hour
Create Disaster	Jonathan	March 5th	March 5th	1 hour	1 hour
Create Terrorism	Lilian	March 14th	March 14th	1 hour	1 hour
Create Event dimension (merging disaster and terrorism)	Lilian, Logan, Jonathan	March 14th	March 16th	1 hour	3 hours
Staging of dimension Country	Logan	March 13th	March 18	1 hour	1 hour
Staging of dimension Month	Logan	March 13th	March 18th	1 hour	1 hour
Staging of dimension Education	Logan, Jonathan	March 13th	March 18	1 hour	1 hour
Staging of dimension Health	Logan, Jonathan	March 13th	March 18	1 hour	1 hour
Staging of dimension Quality of Life	Logan, Jonathan	March 13th	March 18	1 hour	1 hour
Staging of dimension	Logan	March 13th	March 18	1 hour	1 hour

Population					
Staging of Disaster	Jonathan	March 13th	March 18	1 hour	1 hour
Staging of Terrorism	Lilian	March 13th	March 18	1 hour	1 hour
Staging of dimension Event	Lilian, Logan, Jonathan	March 13th	March 18th	1 hour	1 hour
Create HDI measure	Lilian	March 15th	March 15th	1 hour	1 hour
Create Number of Births measure	Lilian	March 15th	March 15th	1 hour	1 hour
Create Number of Deaths measure	Lilian	March 15th	March 15th	1 hour	1 hour
Creating / staging measures	Lilian, Logan	March 15th	March 16th	1 hour	1 hour
Staging of fact table - including FKs and measures	Logan	March 17th	March 18th	2 hours	3 hours
Inputting data into DB	Logan	March 14th	March 18th	3 hours	6 hours
Data quality handling and reporting	Jonathan, Lilian	March 7th	March 17th	3 hours	6 hours
High-Level Plan	Lilian, Jonathan, Logan	Feb 28th	March 19th	1 hour	2 hours
Report documentation	Lilian, Jonathan	March 17th	March 19th	2 hours	2 hours
Meeting Notes	Lilian	February 18th	March 19th	30 mins	30 mins

Meeting Notes

Date: February 28th 2022
Time: 2:30-3 (30mins)
Location: Discord

- Choose 9 countries
- Decided on a conceptual model (professors)

Date: March 5th 2022
Time: 10:30-11:30 (1h)
Location: Discord

- Chose all dimension attributes
- Set up Docker, PostgreSQL

Date: March 7th 2022
Time: 2:30-4:00 (2h30)
Location: Discord

- Create a high level plan
- Update dimensions

Date: March 10th 2022
Time: 2:00-2:45 (45mins)
Location: Discord

- Worked on scripts to grab relevant attributes from data sources for the 9 countries

Date: March 14th 2022
Time: 2:30-3:00 (30mins)
Location: Discord

- Discuss strategies for quality issues

Jonathan:

- Disasters

Lilian:

- HDI
    - Update: No reports were released for 2020
    -

Logan:

- Education

Date: March 15th 2022
Time: 3:00-4:00 (1h)
Location: Discord

- Discuss further strategies/algorithms for quality issues
- Clean data
- Search for additional data sources for events (terrorism)

Date: March 16th 2022
Time: 12:00-1:30 (1h)
Location: Discord

- Clean data, deal with null values/missing data
- Terrorism (either 40 vs 3198 events - by removing minor events)

Date: March 17th 2022
Time: 12:00-1:30 (1h30)
Location: Discord

- Split up remaining work

Jonathan:

- Find new attributes for:
    - Health dimension (need at least 16 attributes)
        - Adults living with HIV
        - Adults newly infected with HIV
        - Children living with HIV
        - Children newly infected with HIV
        - Number of surgical procedures

Check if there's any data for columns:

- Hepatitis_immunization_rate
- dpt_immunization_rate
- measles_immunization_rate
- polio_immunization_rate
- stillbirths
- infant_mortality
- number_of_doctors
- Access_safely_managed_drinking_water
- total_Unemployment_Rate
- Female_Life_Expectancy
- Rural_poverty
- Urban_poverty
- Population_growth

Lilian:

- Merge disasters and terrorism seed into event seed, clean data
- Help check data for columns

Logan:

- Script for fact table

Date: March 18th 2022
Time: 12:00-3:30 (3h30)
Location: Discord

- Continue work and discuss issues

Date: March 19th 2022
Time: 12:00-3:30 (3h30)
Location: Discord

- Finalize codes, search for missing attributes to meet total requirements
- Review high-level plan
- Review git repo
- Finalize report

Github Repository

Link: https://github.com/Logan-Rose/Data-Science-Project