Standard Deviants Data Wrangling Project Proposal

Basic Information

Project Title: *Public Health Database of Children Screened for Malnutrition (in 200 sites around Globe, 21 countries, 12 languages, nearly 180,000 children screened/re-screened)*

Team Members

Name	Email	UID
Bradley Walker	walkerspad@aol.com	u0483666
Monica Regulagadda	u1504835@umail.utah.edu	u1504835
Daniel Goldgar	u0414652@umail.utah.edu	u0414652
Mikayla Compton	u1538218@umail.utah.edu	u1538218
David Bean	u1526388@umail.utah.edu	u1526388

Repository Link: https://github.com/David-Bean/DBMI-Wrangling-Group-3-Project.git

Background and Motivation

There are around 195 million malnourished children globally, with 3 million deaths occurring each year. The prevention and treatment of childhood malnutrition is considered by many to be among the most important and cost-effective charitable efforts. For each dollar invested in this cause, it is estimated that $20 will be generated in the future. Investing can lead to the prevention of death for 7% of children benefited on average, and the avoidance of lifelong physical and cognitive stunting for the rest. This will consequently drastically improve life-long employment and health outcomes. Our team aims to prepare a childhood malnutrition data set for analysis in a hypothetical research project investigating the relationship between economic factors and childhood malnutrition around the world.

Project Objectives

- Categorize children's nutritional status in order to determine needed treatment (calories/micronutrients/family nutrition education) at each of around 200 geographic sites.
- Monitor the selected project implementers (or coordinator performances) in obtaining heights/weights and in influencing cure rates for malnutrition.
- Analyze data and prepare for studies for publication, including secondary data analysis—this has been done without the data wrangling portion just assuming things were OK (but not since 2018).
- Prepare a database to allow for combining screening data and data from WHO Child Growth Database, the World Food Programme Database, and the World Bank Database of social and economic indicators, in order to allow investigators to add many other variables via linking the country or geographic location data from the screening Malnutrition Database with the same information from their Databases.
- Compare outcomes with the cleansed data and non-cleansed data to see what kind of difference the pre-processing work made in the data analysis and outcomes.
- Analyze the effect or correlation of socioeconomic indicators on malnutrition rates seen in the database.
- Develop a "program" to automate a way to pre-process the data in the future.

Data Sourcing

A single project lead in each of the approximate 200 sites is given a tablet with a NoSQL database pre-loaded and receives 4 hours of initial training (additional training available via Zoom from a regional trainer as needed). This project lead then captures the data via in-person-obtained height/weight by two individuals (with a third measure captured in case of significant disagreement between the 1st two measures). A "hard copy" of the data is maintained in-screening location in case of data loss. On obtaining internet access, the project lead pushes a button to "sync" the data to a central location in Utah County. See www.bountifulchildren.org. The database is only available de-identified with permission from Foundation, which was obtained in this instance, and can be seen on the Group 3 Microsoft Teams chat.

Economic data will be primarily sourced from the World Bank Open Data repository at https://data.worldbank.org/**.** Additional sources may be included if necessary.

Data Processing

Data will be placed into a MySQL database and then loaded in Python or R for statistical examination. A Database conceptual design and logical design will be provided. Data will be examined for outliers, nulls, duplicates, and other evaluations suggested by collaborating faculty and/or standard methods used to evaluate data quality as per a selected framework. Based on our preliminary evaluation, we expect significant clean-up. Our evaluation showed that 5% of rows contain outlier(s); and average increases in weight/age data with treatment showed an increase of 5 SDs in improvement from 1st to last measurement. In some cases an entire site may need to have their data removed due to high rates of inconsistencies and/or differences in recording practices.

Quality Visualization

Once attributes of interest have been identified and appropriate cutoff points for what seems physically possible for each category are found, we will design visualizations to show the proportion of entries for each attribute that fall outside of acceptable range. This can be done with aggregative visualizations such as box plots or violin plots for each category, or stacked bar charts to show proportionality. Visualizations may be created to display proportions of imputed data after cleaning. If there are certain data collection locations that have significantly different data from others, we may design visualizations to illustrate those differences and decide whether to exclude them from our analysis.

Must-Have Features

- A combined database consisting of the WHO, WFP, World Bank, and or Foundation datasets.
- Data quality assessment section containing relevant data visualizations
- A cleaned version of the data free from outliers, nulls, duplicates and/or other inconsistencies.

Optional Features

- Automating the process for examining the database performance for future pre-processing.
- Comparison of data outcomes for processed and unprocessed data.

Project Plan and Schedule (tentative)

Objective	Due Date
Finalize Sources	Feb 21
Identify Variables of Interest and Determine cutoff values	Feb 28
Select data quality assessment framework	Mar 07
Instructor Feedback Meeting 1	Mar 07
Design visualizations to illustrate which data that is outside acceptable parameters given by framework	Mar 14
Data quality assessment following appropriate framework	Mar 21
Prepare the data by transforming, aggregating, and filtering as appropriate, based on findings from data quality assessment	Mar 28
Write a final document showing our workflow and demonstrating that our data is prepared for analysis	Apr 4
Instructor Feedback Meeting 2	Apr 4
Create a powerpoint presentation summarizing our work	Apr 11
Project Presentation	Apr 21
Submit Final Document	Apr 22