Group 4: Capstone Kickoff

Aaron Janaszak, Lucas Mueller, Christian Buonfiglio, Addison Gangwish

Variables Used

-Underweight Adults (BMI below 18)

-Water Availability (% people with access to safe drinking water)

-Adult Mortality Rate (chance of death per 1000 people from age 15-60)

-Regions: countries grouped by

Exploratory Questions

-Trends

-Correlations

-Comparisons

-Predictions

Data Platform

-Kafka used to simulate real-time data gathering

-Azure Databricks used for data cleaning and transformation

-Azure Data Lake / Blob Storage used to temporarily store gathered data

-Azure Data Factory used to automate the above process

-SQL database used to store gathered data long-term

-PowerBI used for final visualizations

Data Cleaning

-Filtered data by sex to only be average data between sexes

-Filtered aggregated data (such as “global” or “for a given continent) out

-Used regional table to group countries

-For water and underweight, replace null values with the mean for the country’s region that year

-Filtering data to only use years 2000-2016, since those are shared. For machine learning, will be using entire span for a superior extrapolation.