Group 4 Capstone

Christian Buonfiglio, Addison Gangwish, Aaron Janaszak,
Lucas Mueller

Introduction

Topic: Healthcare

Focus: Health metrics possibly related to mortality rate

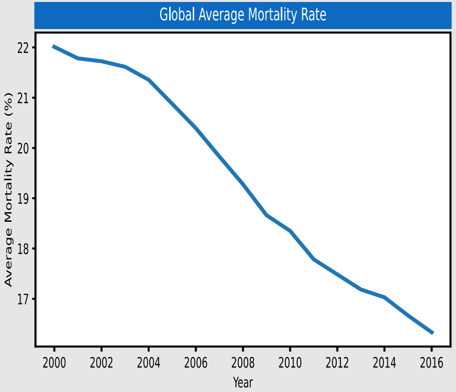
Sources: WHO API, web scraping

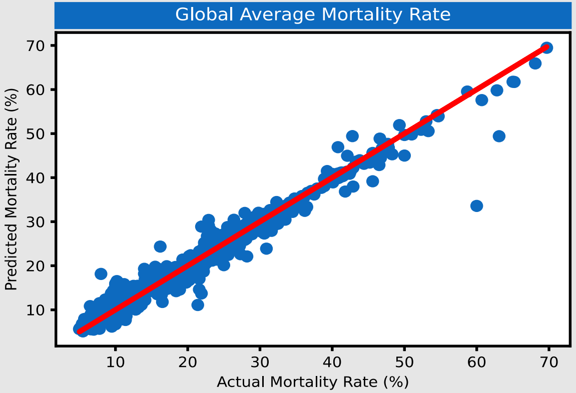
Objective: Create dashboard and ML model to predict mortality rate

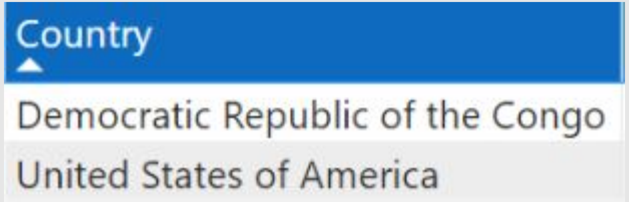
Indicators: Percent of population with access to drinking water, percent of adult population that is underweight (BMI<8), incidence rates of tuberculosis and malaria

Exploratory Questions

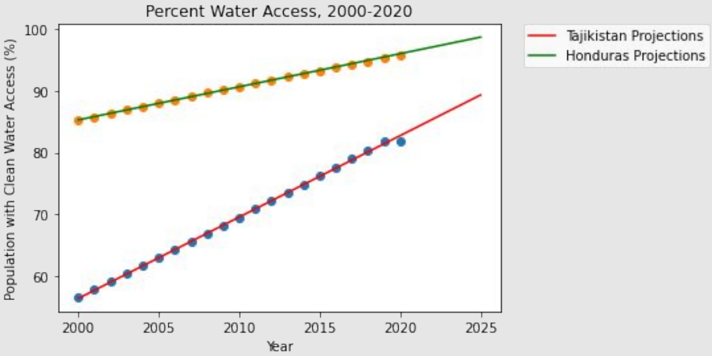
- Trends
- Correlations
- Comparisons
- Predictions



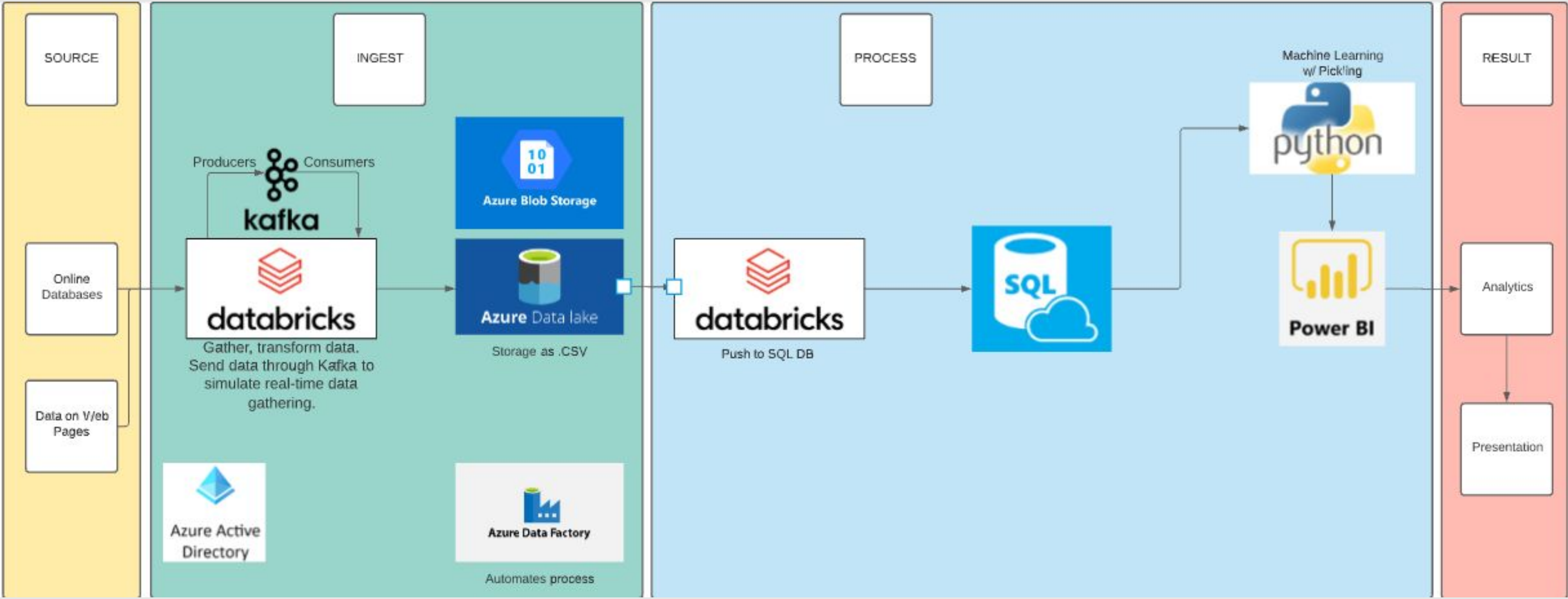




Country ▲
Democratic Republic of the Congo
United States of America



ETL, continued



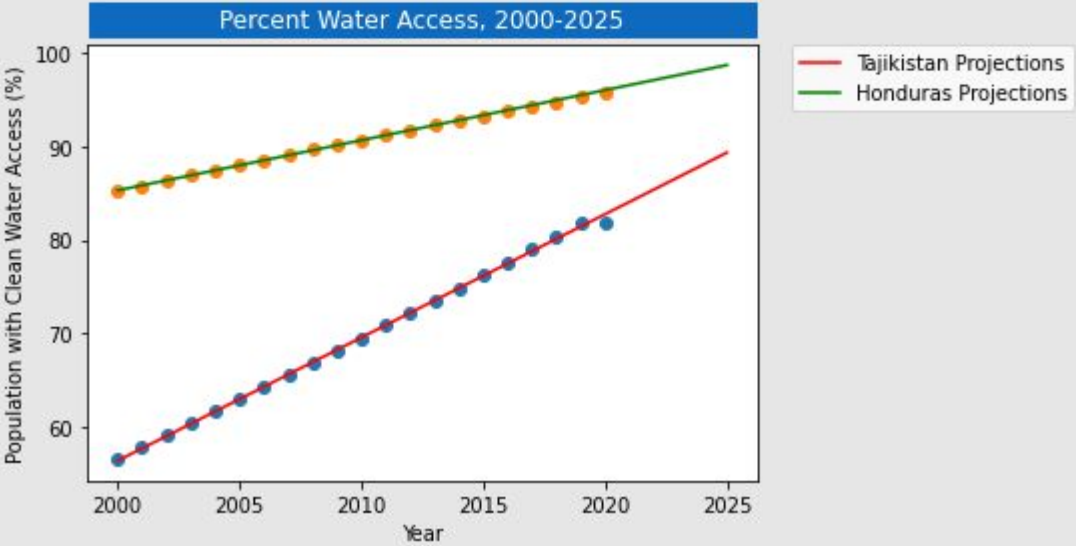
Time Series

Purpose: Linear prediction of future values based on documented trends

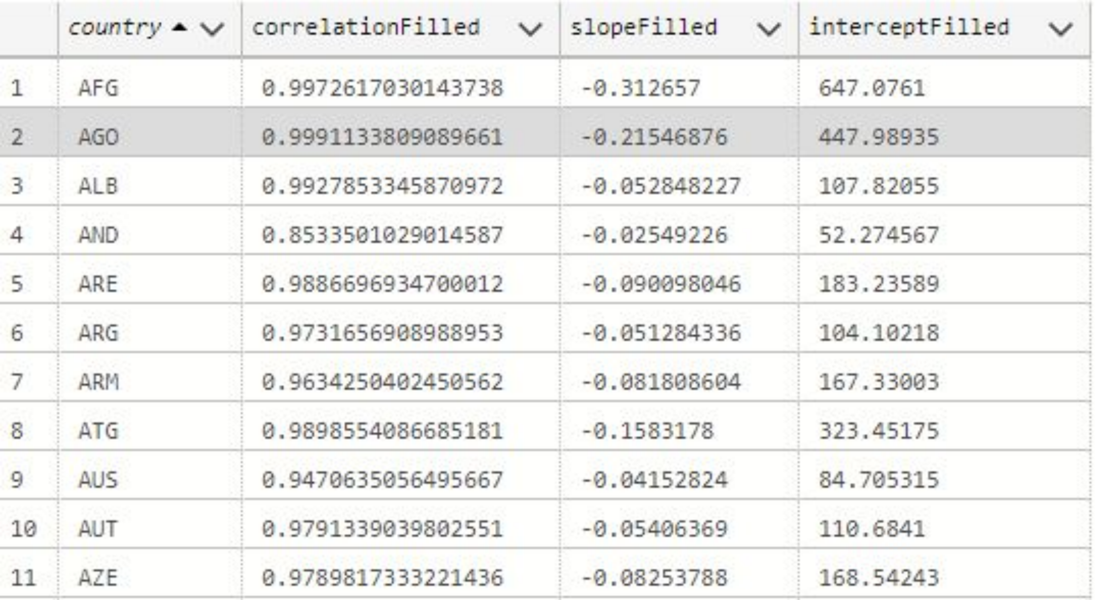
Use these predicted values in ML model to predict mortality rate

The following percentage of countries were at least 85% explained by our linear forecast for each variable:

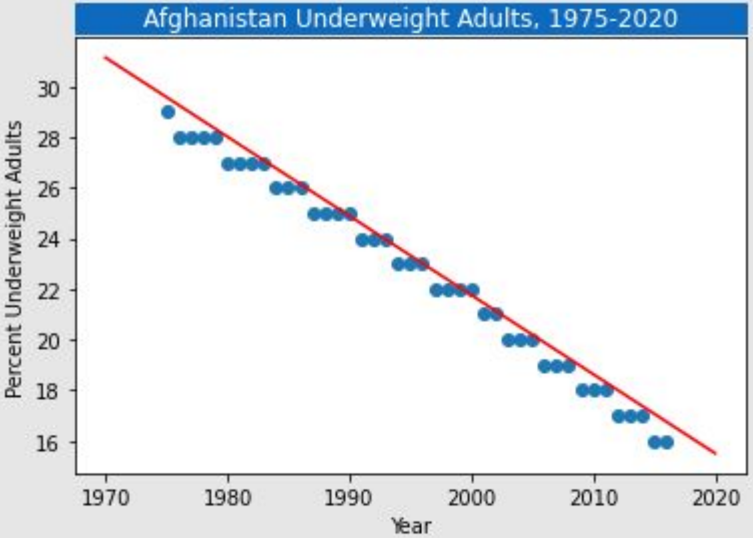
- Underweight: 100%
- Water: 92%
- Mortality: 72%
- Malaria Incidence: 65%
- Tb Incidence: 35%



Time Series Predictions



	country	correlationFilled	slopeFilled	interceptFilled
1	AFG	0.9972617030143738	-0.312657	647.0761
2	AGO	0.9991133809089661	-0.21546876	447.98935
3	ALB	0.9927853345870972	-0.052848227	107.82055
4	AND	0.853350129014587	-0.02549226	52.274567
5	ARE	0.9886696934700012	-0.090098046	183.23589
6	ARG	0.973165690898953	-0.051284336	104.10218
7	ARM	0.9634250402450562	-0.081808604	167.33003
8	ATG	0.989855408668518	-0.1583178	323.45175
9	AUS	0.9470635056495667	-0.04152824	84.705315
10	AUT	0.979133903980255	-0.05406369	110.6841
11	AZE	0.9789817333221436	-0.08253788	168.54243

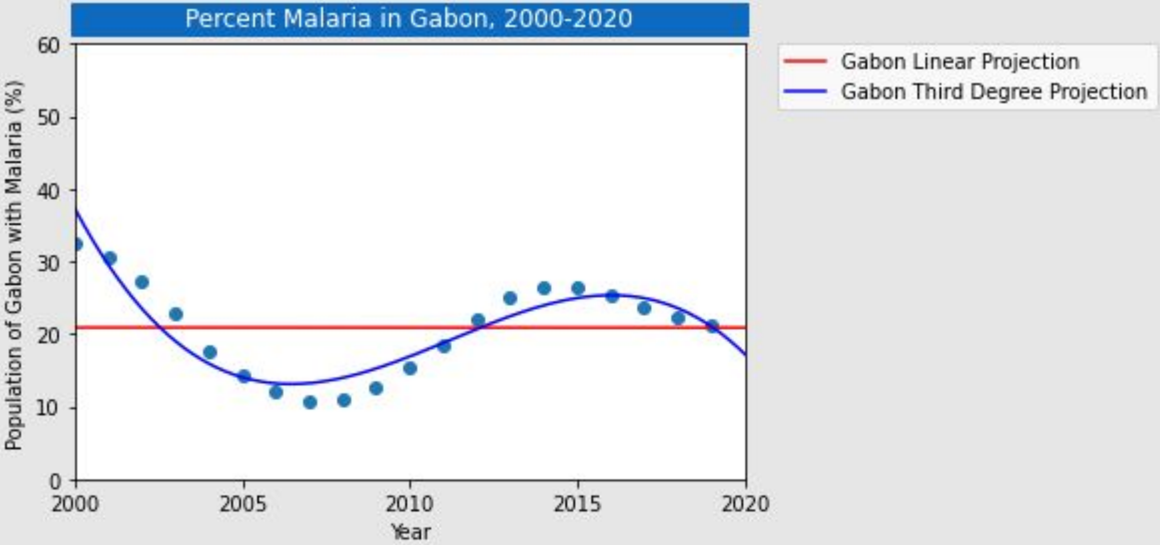


Time Series Predictions

-Not all variables were linear

-Disease is essentially impossible to model given the paucity of data.

-Crude polynomial fits worked but were almost certainly overfitting to the data.



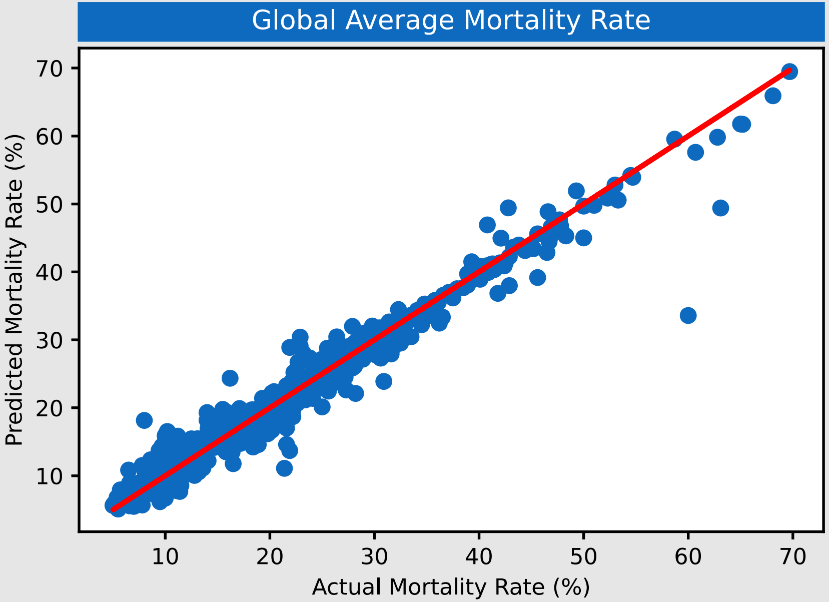
ML Model

Attempts:

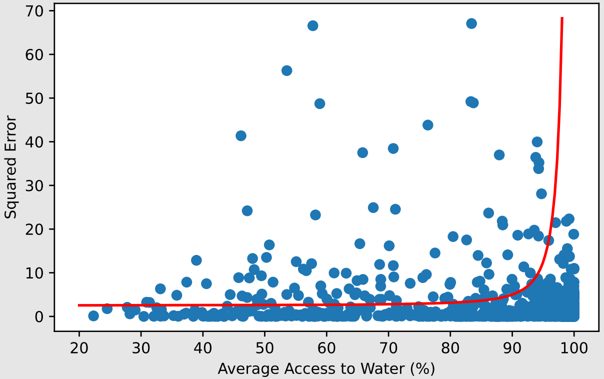
- Basic Multilinear Regression
- Generalized Linear Regression
- Basic Decision Tree
- GBT Regressor
- K-Nearest Neighbors
- Radius-Nearest Neighbor
- MLP Regressor
- Gaussian Process Regressor
- Decision Tree Regression
- Random Forest
- Gradient Boosting Regressor
- Ada Boosting Regressor

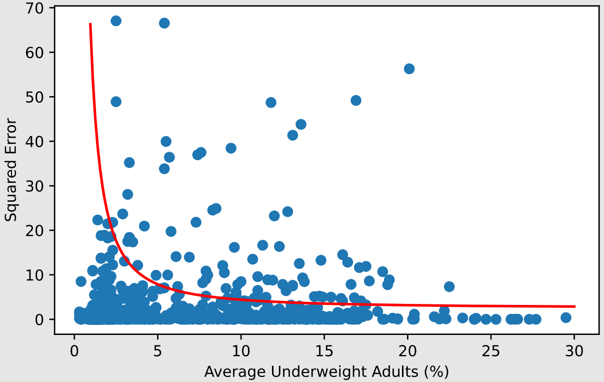
Best Model: Extra Trees Forest

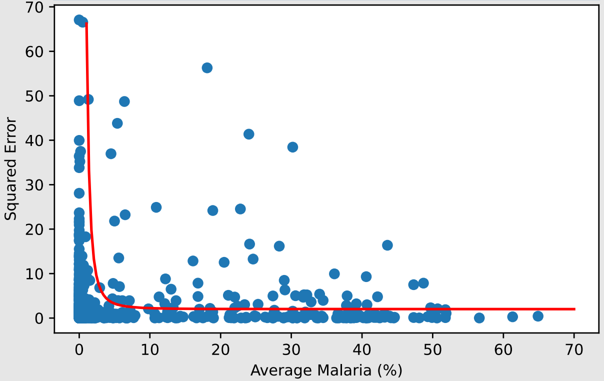
Results: 0.97

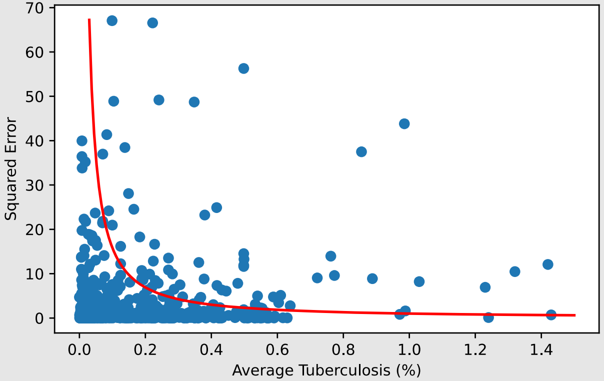


ML Model


Squared Error vs Access to Water (%)

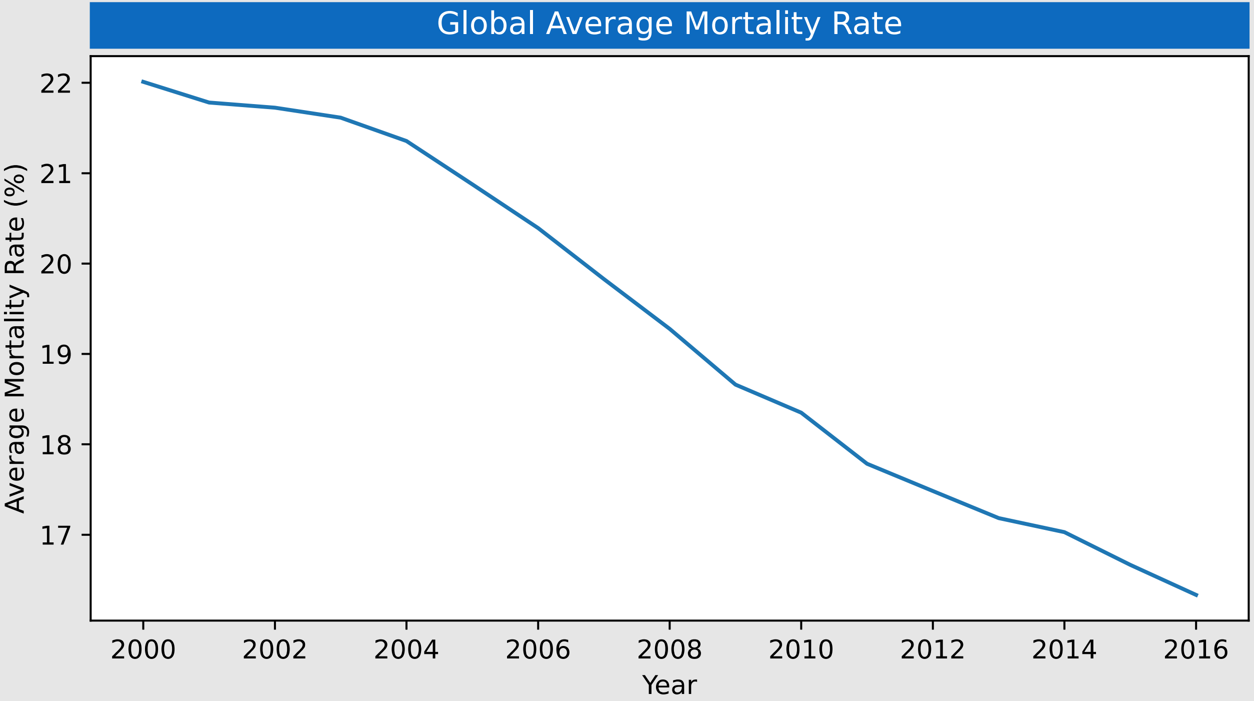

Squared Error vs Average Underweight Adults (%)


Squared Error vs Average Malaria (%)


Squared Error vs Average Tuberculosis (%)

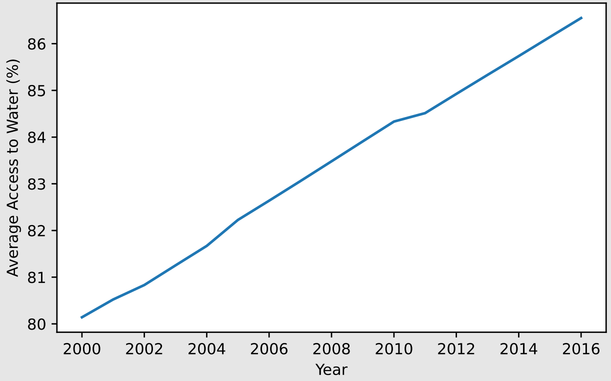
Major Findings

Global average mortality rate dropped from ~22% to ~16% from 2000 to 2016

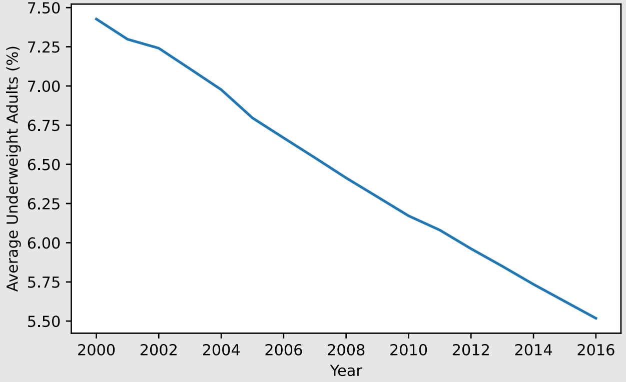


Global Trends

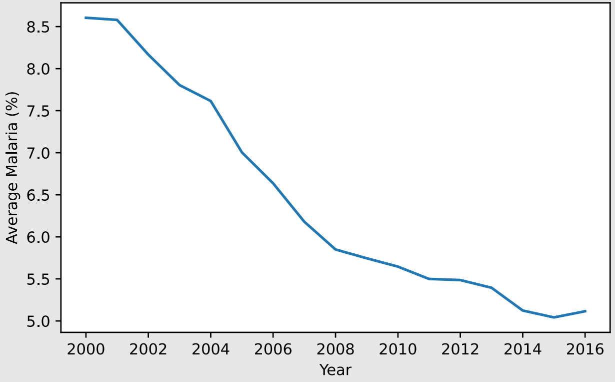
Global Average Access to Water (%)



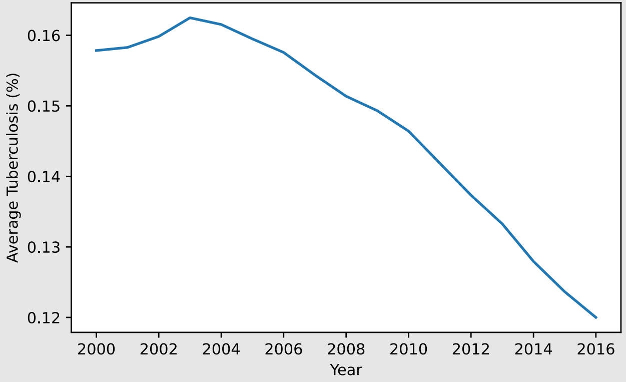
Global Average Underweight Adults (%)



Global Average Malaria (%)

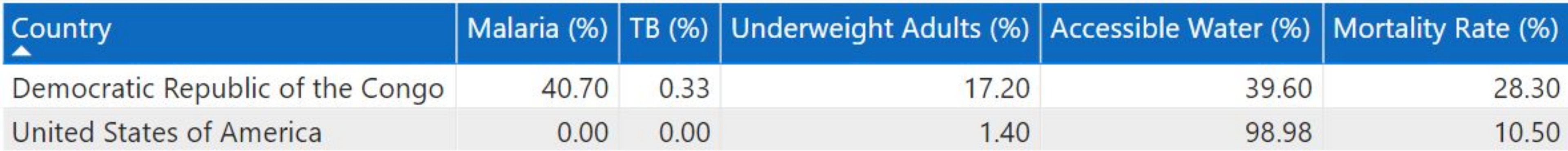


Global Average Tuberculosis (%)

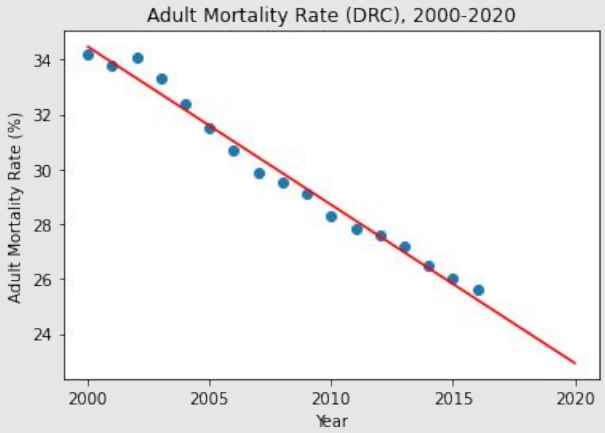


Major Findings

The Midwest portion of the United States had roughly the same population as the Democratic Republic of the Congo in 2010.



Country	Malaria (%)	TB (%)	Underweight Adults (%)	Accessible Water (%)	Mortality Rate (%)
Democratic Republic of the Congo	40.70	0.33	17.20	39.60	28.30
United States of America	0.00	0.00	1.40	98.98	10.50



MIDWEST
ND
MN
SD
WEST NORTH CENTRAL
WI
MI
EAST NORTH CENTRAL
IA
NE
IL
IN
OH
KS
MO
KY
W

Recommendations and Conclusion

Improving and maintaining water infrastructure will be vital going forward

In general, all statistics are improving

We could use more indicators (HIV, ischemic heart disease, etc.) and/or breakdown into smaller, specific groups by age or sex

Thank you