Time Series Analysis of Global Health Trends

Megan Eddy

Bellevue University

Business Problem

This project will explore these trends over time and correlations before ultimately predicting values over time.

Background

“Global Health has become one of the most important areas of foreign, development and security policy in the past 15 years” (Holst, 2020). It stands to reason that as medical care evolves, the overall health of the world should improve. But is this the case and if so, how can these changes be predicted? Are there telling correlations between certain health indicators?

Dataset

Three datasets were included in the download. The datasets were sourced from the World Bank Open Data portal by the user and published to Kaggle. This source is open to the public. The version of the data published to Kaggle was last updated on 12/26/2024 as noted on the datasets themselves.

The largest file includes indicators of global health trends, sorted by year and country. The second dataset indicates each country’s geographic location as well as what income category it belongs to. Additionally, there are notes outlining ways data has been adjusted for specific countries. The third dataset gives information about how the measurements for each of the indicators were collected.

Methodology

The dataset containing global health trend values was merged with the second dataset to allow for division of individual countries into regions. The dataset was restructured, and EDA was employed to identify an appropriate target variable. After the target variable was selected, a baseline model was built using ARIMA to model the target variable, and evaluation metrics were used to measure its effectiveness. PCA was performed to handle multicollinearity and a SARIMAX model was built. Results from this were evaluated against the R-squared, RSME and MAE values of the baseline model.

Analysis

The dataset was merged and pivoted, setting the time values (Year) as the index and each indicator was recorded in individual columns.

EDA was employed to identify an appropriate target variable. Missing values were reviewed by indicators, regions and years. One region value (World) was selected for model building and the target variable chosen was indicator SP.DYN.LE00.FE.IN (female life expectancy at birth).

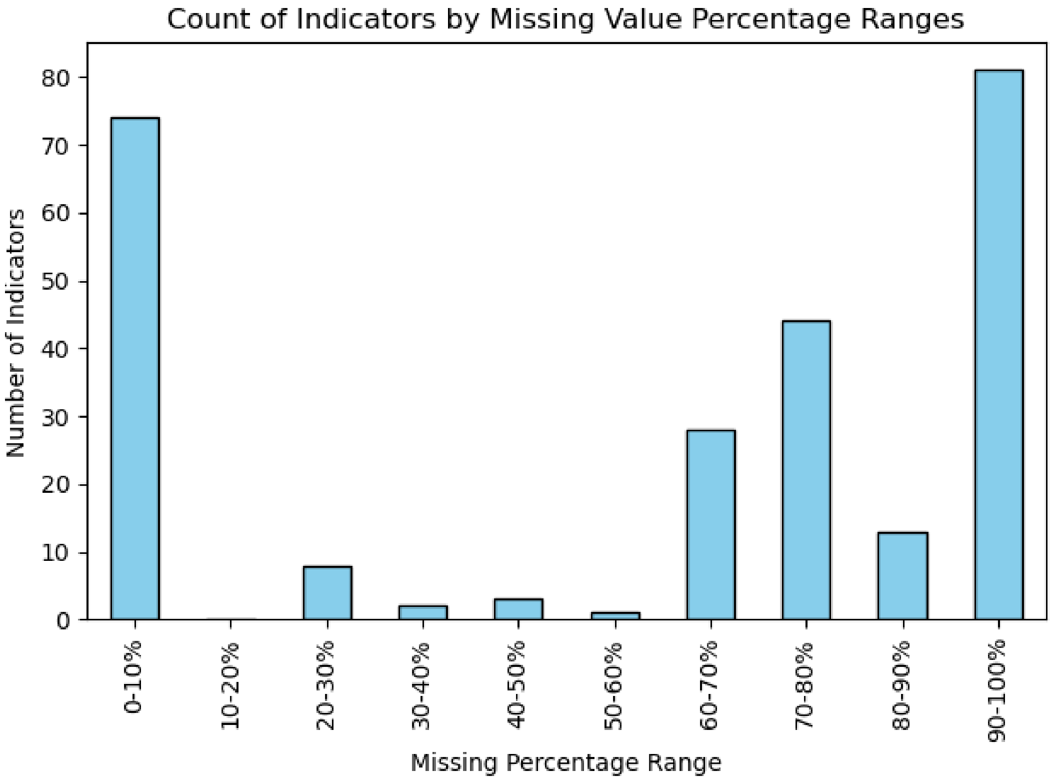

Figure 1: The Percentage of Missing Values by Health Indicator

ADF values were evaluated to check for stationarity, variance was tested and a correlation matrix built to verify the target variable’s suitability.

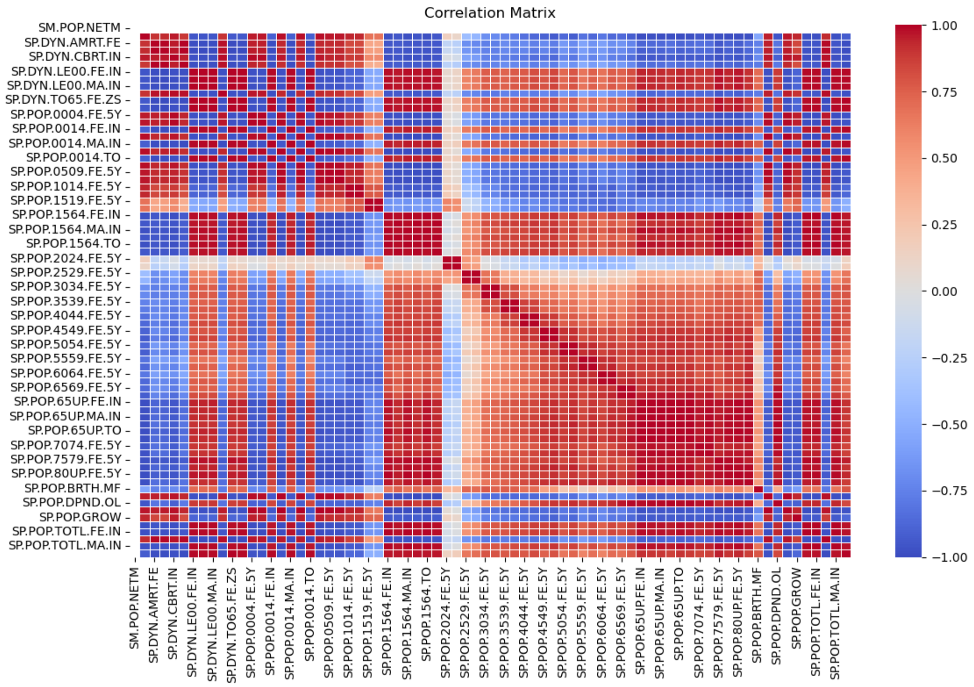

Figure 2: Correlation Heatmap of Filtered Indicators

ADF Statistic of Target Variable: SP.DYN.LE00.FE.IN

```
ADF Statistic: -3.3126022576296115
p-value: 0.014324465705779778
Critical Values:
1%: -3.548493559596539
5%: -2.912836594776334
10%: -2.594129155766944
Data is stationary
```

Figure 2: ADF Statistic of Target Variable

Multicollinearity between variables was managed using VIF outputs and PCA. The PCA components were merged with the target variable and time series values. The target variable was found to be non-stationary at that time, so differencing was performed twice. A SARIMAX model was built and forecasts predicted.

The baseline model performed better than the SARIMAX model with the added predictors. Neither model performed better than guessing the mean value.

Conclusion

By paring down the large dataset to one region, missing values and computational needs for model building were made more manageable. The target variable was originally found to be stationary but showed a strong deterministic positive trend that caused issues with the model. The residuals showed no autocorrelation but were not normally distributed. The variance of errors was not constant.

Another modeling type such as Exponential Smoothing or XGBoost might be a more suitable choice to model this data.

Assumptions

The assumption is being made that a model built to evaluate the life expectancy of females at birth for the World region would also be useful in predicting this same variable across all regions.

Limitations

Perceived strain on computational resources led to paring down of the model being built. Currently the model will only represent the outcome for the World region.

Time was a limitation as more data cleaning and verification is needed to create a working model.

Challenges

Missing data played a large role in the selection of the region and the target variable. Imputation was likely possible on a broader scale, but each region or possibly each country would need to be evaluated to most accurately determine the appropriate method. This prevented a choice of a more interesting target variable.

Feature selection was also challenging as there were many redundant variables. For example: male population, female population and total population.

Future Uses/Additional Applications

This model could be used to identify other indicators of global health or for individual countries or regions. It could assist in targeting specific types of aid for government or health organizations.

Recommendations

Imputation of missing values is recommended to make the model more robust. This would need to be done on a granular level so values for countries are not skewed by other countries.

Reviewing the model selection and trying another model choice would likely improve the results. Handling multicollinearity in another manner might also help to develop a working model with the data.

Implementation

To implement this model, the baseline model will need to be built and evaluated. Then feature selection will be performed using PCA and removal of fields at risk of multicollinearity. An ARIMA model will then be built using the selected features to aid in the prediction of the target variable.

Ethical Considerations

Currently the model only considers the World region. This means that results are general and do not specifically apply to any single country. Life expectancy rates vary by country so predicted outputs should be seen through the global lens. The model is also currently only using one indicator of global health. As shown in the dataset, there are many others that could also be considered. Female life expectancy at birth is not necessarily the best nor most predictive of the overall health of any given population.

References

Holst, J. Global Health – emergence, hegemonic trends and biomedical reductionism. *Global Health* 16, 42 (2020). https://doi.org/10.1186/s12992-020-00573-4

Jatt, H. (2025). World Health Indicators Dataset [Dataset]. Kaggle. Retrieved from https://www.kaggle.com/datasets/realhamzanet/world-health-indicators-dataset?select=API_8_DS2_en_csv_v2_3654.csv

Kirwan, D. (2009). Global health: Current issues, future trends, and foreign policy. *Clinical Medicine, 9*(3), 247-253. https://doi.org/10.7861/clinmedicine.9-3-247