Project Report

GitHub URL

https://github.com/aregan1/UCDPA_amieregan

Abstract

This project aims to show the correlation of life expectancy at birth with the availability of home hand washing facilities. It features two complementary data sets showing WHO Life expectancy at birth for 66 countries and the availability of home hand washing facilities in those countries. A number of variables are included in the data including, time period (annual), gender and urban/rural settings.

Introduction

This dataset contains health statistics measured by the WHO. The owner of this dataset has provided 39 separate datasets; through filtering of the data to reflect various subcategories to provide more manageable datasets for analysis. The data is comprehensive, covering the time period 2000-2019 worldwide.

I chose this dataset as it has relevance to my work at the Health Research Board, where I manage funding for a number of clinical trials and interventions in the Covid-19 area. We know that healthy life expectancy will have been significantly impact by the current pandemic. Recent studies have shown that more than 28 million excess years of life were lost in 2020 in 31 countries worldwide. This rate of loss is higher in men than in women1[iii].

I have used this exercise to familiarise myself with WHO healthy life expectancy figures when compared to other factors (in this case the availability of handwashing facilities), so that I will be able to compare other factors to the updated WHO figures for 2020 onwards. In particular I would be interested in assessing the reduction in life expectancy when compared to investment in health services research and clinical trials on a country-by-country basis. This information is not yet available in a usable format.

Dataset

I sourced this dataset from Kaggle (https://www.kaggle.com/utkarshxy/who-worldhealth-statistics-2020-complete).

This dataset contained 39 separate datasets which provided information under the following categories:

- Life expectancy and Healthy life expectancy
- Maternal mortality
- New-born and child mortality
- Communicable Diseases
- Noncommunicable diseases and mental health
- Substance abuse

- Road Traffic Injuries
- Sexual and reproductive health
- Achieve universal health coverage (UHC) including financial risk protection
- Tobacco control
- Health Workforce
- Eliminate violence Against women and girls
- Drinking Water/Sanitisation
- Clean household energy

From the 39 available datasets I chose the following for analysis as part of this project:

1. lifeExpectancyAtBirth.csv -> Life expectancy at birth, country wise mentioned in age (years).
2. basicHandWashing.csv -> Population with basic handwashing facilities at home (%)

I chose this data as I found the datasets provided to be manageable and of good quality. The information provided was in a mix of formats both integers and strings, which required almost no clean up. The data was also split in such a way as to allow for easy demonstration of the merging of datasets. Importantly the two datasets contained data from overlapping time periods allowing for direct comparison of indicators. The data was also from a reliable primary source (WHO) and so I was confident in its quality and accuracy.

Implementation Process

1. Selection and import of datasets:
   lifeExpectancyAtBirth.csv -> Life expectancy at birth, country wise mentioned in age (years).
   basicHandWashing.csv -> Population with basic handwashing facilities at home (%)
2. Renaming of datasets to hand_wash_2010 and Life_expectancy_2010.
3. Interrogation of dataset quality for clean-up- no clean up required. The datasets contained no blanks or null values.
4. Sorting of data to see what period is common to both datasets (2015 and 2010).
5. Create a mask to only show data for 2010 in both datasets, I chose to use mask over sorting as it would create a new dataset which could be used in further work reducing the risk of my selection of the wrong dataset. Given the size of the hand_wash dataframe I felt this was a better approach as I wouldn't have to check the dataframe line by line to make sure the sorting was in place each time.
6. Removal of unnecessary columns in both datasets such as indicator.
7. Remove Urban and Rural values from hand_wash_2010 to only show total % of households with adequate handwashing facilities. This allows for direct comparison of this value with life expectancy at birth.
8. Merging of datasets: hand_wash_2010 and Life_expectancy_2010.
9. Removal of columns which were not required for analysis.
10. Demonstration of dictionary and lists using original csv (lifeExpectancyAtBirth.csv). I had some issues here trying to carry this out in the merged dataframe which carried over into the data visualisation below.
11. I was unable to covert my merged dataframe to a NumPy array- I believe this was due to the different data types; sting and integers- I was not able to solve this problem…
12. Data Visualisation: use of matplotlib to plot the following:
    a) Scatter plot displaying the correlation between life expectancy at birth and the availability of adequate handwashing facilities in the household. Each data point on the

plot represents a country. A scatter plot was the best choice in this case as the data to be visualised consists of two paired data points whose relationship was to be interrogated for a positive or negative correlation.

b) Bar chart showing life expectancy at birth (in years) in ascending order for each country within the dataset. In this case I chose a bar chart as there was only one variable, from one time point to be charted (life expectancy) but where we wanted to show a relationship between these figures (country by country).

c) Bar chart showing the availability of adequate handwash facilities (by % of households) in ascending order for each country within the dataset. As above I chose a bar chart as this was one data point taken at a single time point. The bar chart allows for direct comparison of handwashing facilities on a country-by-country basis.

Results

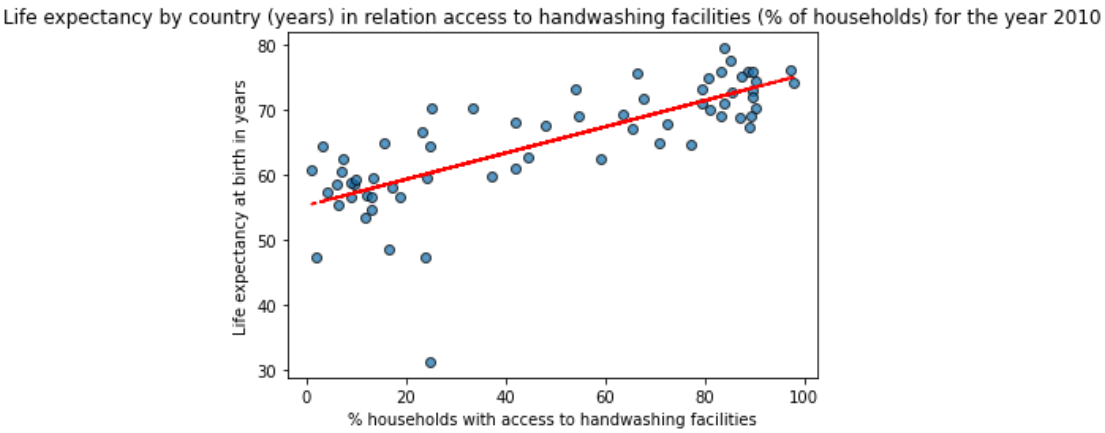

Fig.1 Scatter plot displaying the correlation between life expectancy at birth and the availability of adequate handwashing facilities in the household. This data relates to the year 2010 only as reported by the WHO. Each point represents the data for a single country.

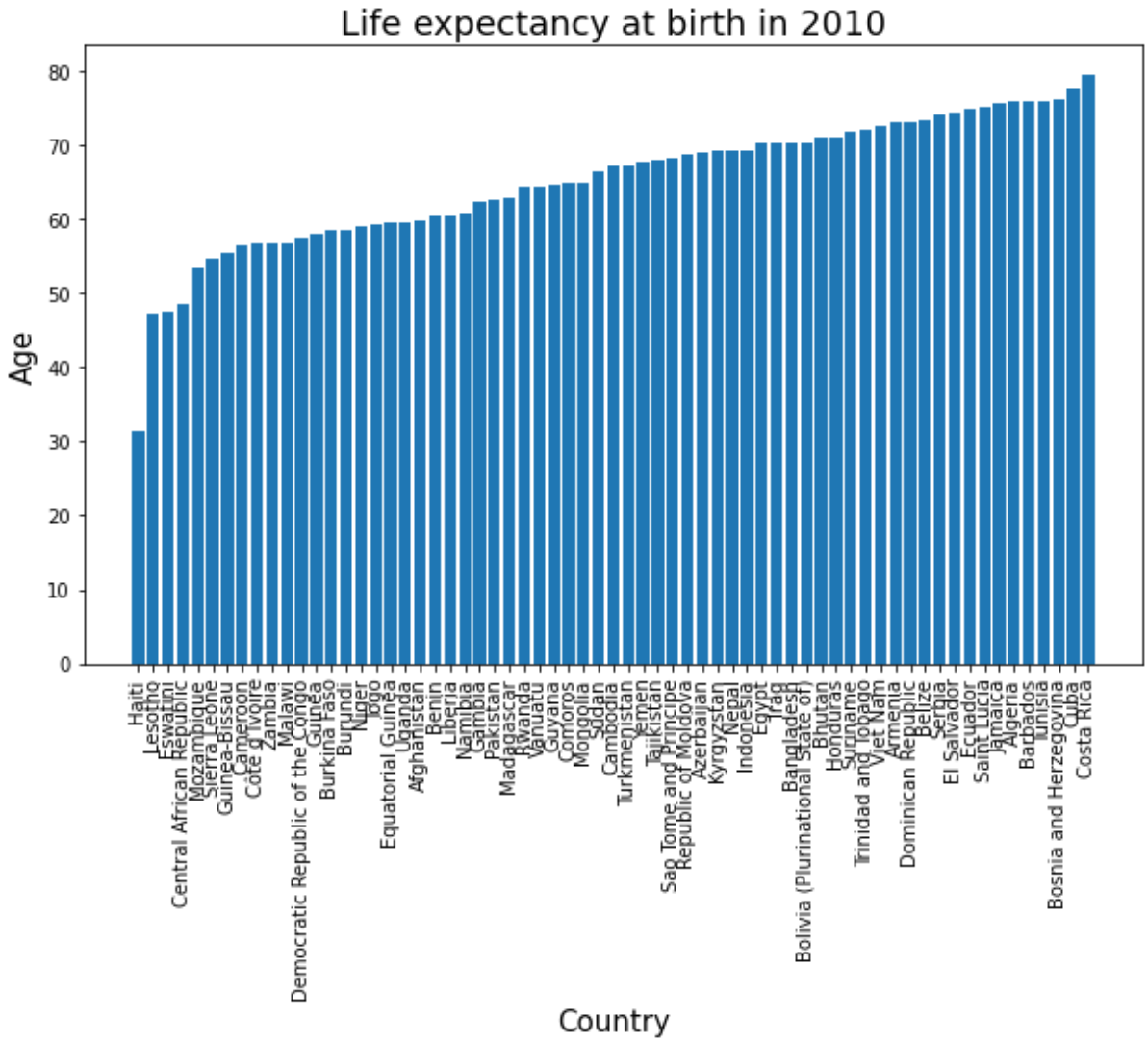

Figure.2. Bar chart showing life expectancy at birth (in years) in ascending order for each country within the dataset. This data relates to the year 2010 only as reported by the WHO.

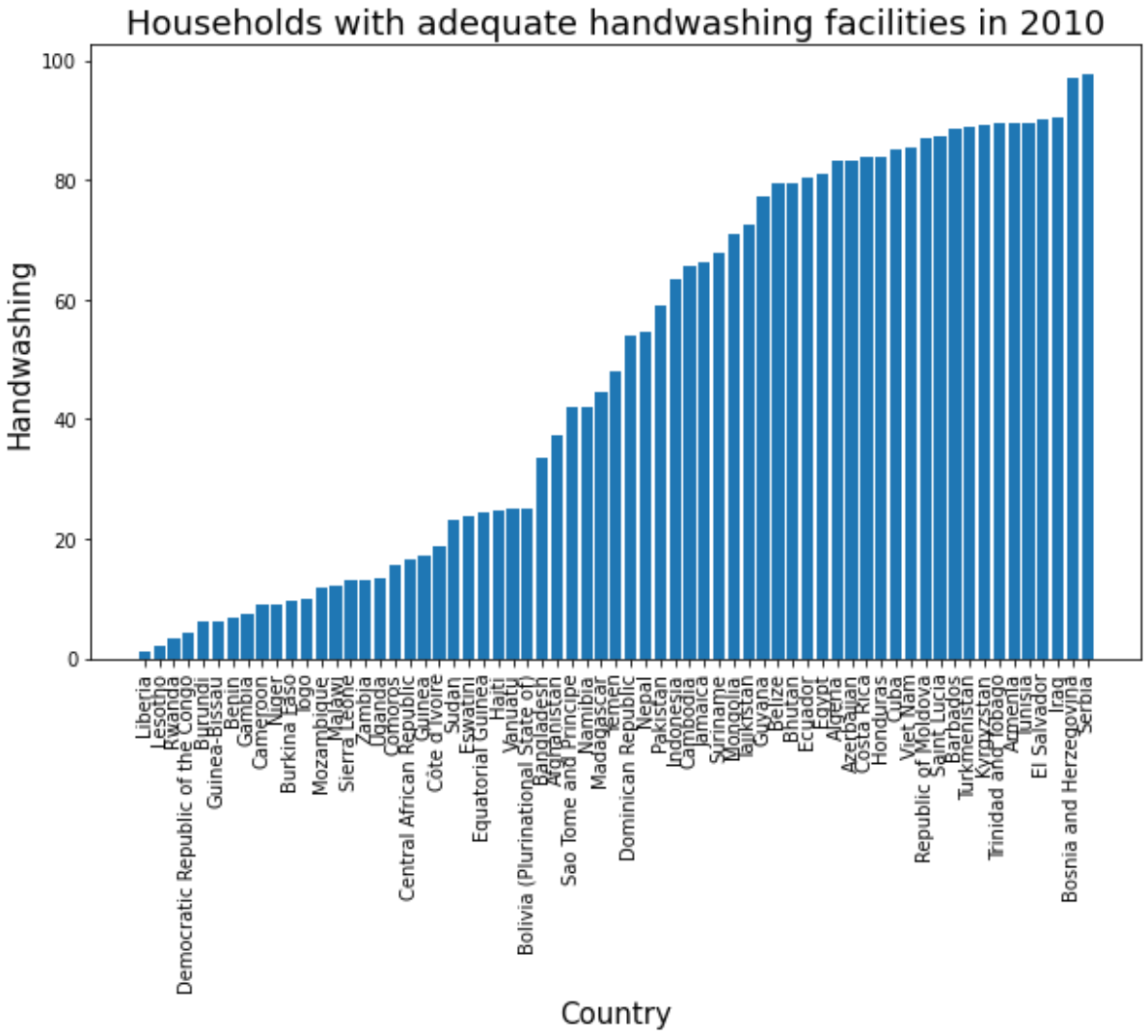

Figure.3. Bar chart showing the availability of adequate handwash facilities (by % of households) in ascending order for each country within the dataset. This data relates to the year 2010 only as reported by the WHO.

Insights

- From the scatter plot we can see that there is a strong, positive, linear correlation between the availability of adequate handwashing facilities in a household and the Life expectancy at birth for a person in that household. This pattern is seen across 64 different countries for the year 2010. Using the trend line we can see that there is good grouping of data points however there is one outlier which I will address in further comments below.
- From the bar charts we can see that the highest life expectancy at birth for the year 2010 is seen in Costa Rica. However Costa Rica has only the 16th highest % of household with adequate handwashing facilities indicating that there are other variables which influenced the life expectancy at birth figures for Costa Ricans. Inclusion of further datasets to explore other variables might give us insight into this.

- Liberia appears to have the lowest % of households with access to adequate handwashing facilities (1.15%), however this does not result in Liberia having the lowest life expectancy at birth for the year 2010 (60.7 years). In 2010 the country with the lowest life expectancy at birth was Haiti at 31.28 years (the outlying data point seen in the scatter plot). The limitations of this type of analysis is clear- without further information the data for Haiti seems anomalous. The raw data does not control for a mass fatality event which befell Haiti in 2010. A catastrophic magnitude 7.0 earthquake struck Haiti at 16:53 local time (21:53 UTC) on Tuesday, 12 January 2010[iii].
- Although there is a positive linear correlation between life expectancy and handwashing (as shown by the scatter plot), discrepancies on a country-by-country basis (shown by the bar charts) indicate that in this case correlation is not causation.
- All of the countries listed within this dataset are in the developing world. This was due to filtering of the datasets to show only the year 2010, the year for which data was recorded in both datasets (handwashing and life expectancy). We therefore cannot say that this is an exhaustive study of the relationship between handwashing facilities and life expectancy as the date is heavily skewed towards the developing world where we would expect to see lower life expectancy and reduced handwashing facilities. An analysis including countries from the developed world would be much more representative of the relationship being examined here.

References

[i] Islam N, Jdanov D A, Shkolnikov V M, Khunti K, Kawachi I, White M et al. Effects of covid-19 pandemic on life expectancy and premature mortality in 2020: time series analysis in 37 countries BMJ 2021; 375 :e066768 doi:10.1136/bmj-2021-066768

[iii] https://en.wikipedia.org/wiki/2010_Haiti_earthquake