# Indicator Analysis & Decisions

## 1. Connectivity Indicators

### 1.1 Internet Users & Cellular Subscriptions

**(No missing values)**: The dataset was formed using a weighted average and should therefore be used with caution. However, as far as we can see, the original source of data for both Internet Users and Cellular Subscriptions correlates with our data.

→ *Decision: Keep*

The issue with a weighted average aggregation method is that it only allows for country comparison over a period of time (from 1960 - 2020). For example, in terms of internet users for Iraq (an outlier):

- Our WA data states that 1% of the population is using the internet.
- The original data states that 75% of the population in 2017 are using the internet.

The same can be said of cellular subscriptions in the case of Kuwait (an outlier):

- Our WA data states that 57% of the population has cellular subscriptions.
- The original data states that 98% of the population in 2017 has cellular subscriptions.

Figure 1:
Distribution of Internet Users by Income Group

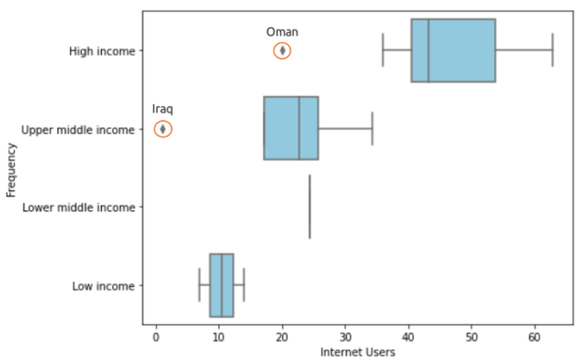
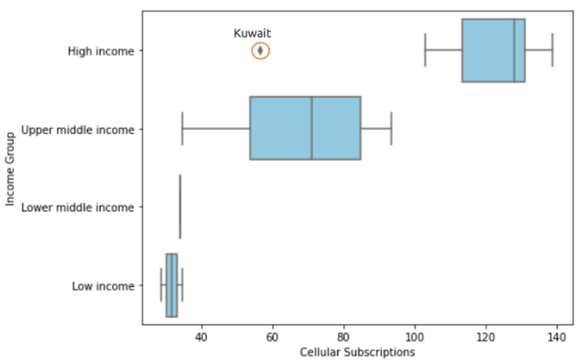



Figure 2:
Distribution of Cellular Subscriptions by Income Group




## 2. Economic Indicators

As in figure 3, there are seven economic indicators. While Syria, Yemen and West Bank and Gaza experience civil war situations (CIA, 2020a)( CIA, 2020b)( CIA, 2020c)( CIA, 2020d ), the economic situation is challenging resulting in a low to lower middle income level as can be withdrawn from figure 3. The other countries in the region have upper middle to high income levels and hence, assumingly a considerably stable economy.

| Country | Income Group |
|---|---|
| Syrian Arab Republic | Low income |
| Yemen, Rep. | Low income |
| West Bank and Gaza | Lower middle income |
| Iraq | Upper middle income |
| Jordan | Upper middle income |
| Lebanon | Upper middle income |
| Turkey | Upper middle income |
| United Arab Emirates | High income |
| Bahrain | High income |
| Cyprus | High income |
| Israel | High income |
| Kuwait | High income |
| Oman | High income |
| Qatar | High income |
| Saudi Arabia | High income |

*Figure 3: Income Group By Country in the Arabian Peninsula*

### 2.1 GDP per person employed

**(6.67% missing values)**: GDP per employed person represents the labor productivity in each country and is estimated according to national account conventions to allow comparisons. These national account conventions differ significantly in their topicality: While Iraq, Jordan and Syria use systems from 1968, others updated their system at 1993 and some other countries are using a system from 2008. West Bank and Gaza does not even have a system for national accounts, which explains why it is the only country missing a value for this indicator. Furthermore, the metadata hints "there are still significant limitations on the availability of reliable data".

→ *Decision: Drop*

### 2.2 GNI per capita

**(6.67% missing values)**: As of the metadata, GNI per capita is computed using World bank's National Accounts data and OECD National Account data. The combination of both data sources gives confidence in the reliability, and the source data matches the given data set.

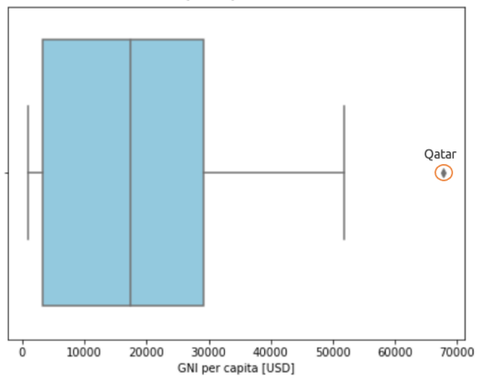
→ *Decision: Keep*

The only country without this indicator is Syria (SYR). CIA (2020a) states that Syria's economy is deteriorated because of the civil war and ongoing humanitarian crisis within the country. The missing value will be imputed with the mean of West Bank and Gaza's and Yemen's GNI per capita because those two countries face similar economic conditions. The imputed distribution is not deviating much from the original distribution as of figure 4.

As visualized in figure 5, Qatar is an upper outlier in the region. The massive oil and natural gas reserves in the country are making it one of the richest countries in the world and leading to the highest GNI per capita in the region(CIA, 2020e).

Figure 4:
GNI per Capita Distribution
(With/without imputation of missing values)


Figure 5:
GNI per Capita Outliers



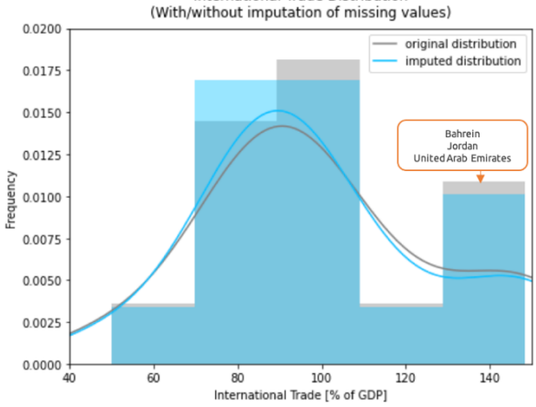

## 2.3 International Trade

**(6.67% missing values)**: As of the meta data, International Trade is the sum of import and exports measured as a share of GDP. It is computed using the same data sources as GNI per capita.

→ *Decision: Keep*

As before Syria is missing a value for this indicator. Again, this will be imputed with the mean of West Bank and Gaza's and Yemen's value due to the similar economic situations of the countries. The imputed distribution is not deviating much from the original distribution as of figure 6. There are no outliers for this indicator according to the box plot. However, it is noticeable that Bahrain, Jordan and ARE are not part of the statistically expected normal distribution as they have indicator values above 130 % of GDP.

Figure 6:
International Trade Distribution
(With/without imputation of missing values)



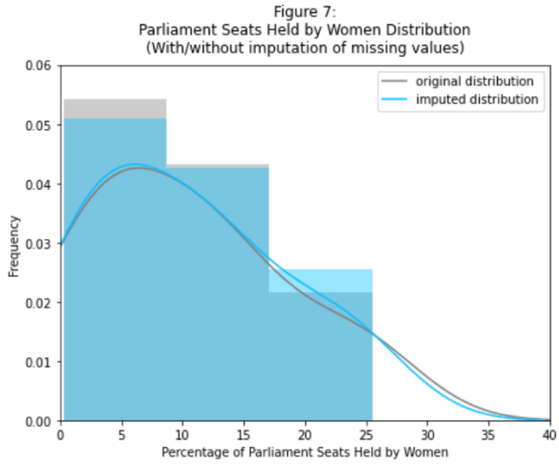

## 2.4 Parliament seats hold by women

**(26.67% missing values)**: This indicator gives a hint on the gender equality in the region. Obtaining this information can be difficult with ad hoc changes like replacements due to death or resignations and the reliability can be questioned because it is compiled on national information. However, the numbers from the dataset and the source seem to be aligned with the Political Empowerment Section of the Global Gender Gap Report (World Economic Forum, 2020). Furthermore, it is a Sustainable Development Goal Indicator (United Nations, 2020).

→ *Decision: Keep*

Oman, Qatar and Saudi Arabia are missing values for this indicator while the data source actually provides data for them (IPU, 2020a)(IPU, 2020b)(IPU, 2020c). These data points will be used to impute the missing values.

Figure 7:
Parliament Seats Held by Women Distribution
(With/without imputation of missing values)




As the indicator is only compiled for countries with an existing national legislature, there is no data available for West Bank and Gaza. The Palestinian Legislative Council was dissolved in December 2018 and since then t there is no parliament in place (CIA, 2020c). Hence, it is decided to not impute the null value for PSE.

The imputed distribution is not deviating much from the original distribution as of figure 7 and there are no outliers.

## 2.5 ODA per Capita¶

**(46.67% missing values)**: Official Development Assistance (ODA) can be received by countries that are on the DAC list of aid recipients (OECD, 2020). This indicator displays to what extent a country received ODA. As it does not take into account how the recipient countries gave much aid to other developing countries, some countries might be reflected as aid receivers while they are actually net donors. Due to the missing counterpart, this indicator can be misleading and is not seen as meaningful.

→ *Decision: Drop*

## 2.6 BOP income share

**(80% missing values)**: Poverty data like the income share held by the lowest 20% of the population are difficult to obtain. The indicator is retrieved from national household surveys, which are typically computed every few years. Often, the insights from the household survey are not comparable due to different computation methods and times. Furthermore, it is only filled by three countries: Cyprus, Jordan, and Turkey. These have a strong affiliation with the European Union by being a member state, candidate, or a close relationship (European Commission, 2019).

→ *Decision: Drop*

## 2.7 Poverty Line Gap

**(93.33% missing values)**: As BOP income, this indicator is retrieved from national household surveys and faces the same difficulties. In our region, only Jordan is reporting the indicator, while worldwide, over 90% of the countries do not report it. Jordan has implemented a poverty reduction strategy, which is the reason for them monitoring the poverty line gap closely (UNDP, 2013

→ *Decision: Drop*

# 3. Education Indicators

## 3.1 Enrollment Rate

- Adjusted net enrollment rate, primary
- School enrollment, primary

**(46.67% missing values)**: These indicators are considered as a pair due to an identical definition and source; also, the data points for each enrollment rate differ in amount by an average of 1.5. Furthermore, 40% of values for all regions is null. Secondly, the datasets source only provides "Adjusted net enrollment rate, *one year before the official primary entry age*." But according to our metadata, the enrollment rates are the percentage of children *within the school-age group for primary education*. Therefore, the data cannot be verified, rendering it untrustworthy.

→ *Decision: Drop*

## 3.2 Completion Rate

**(40% missing values)**: The data from the original source contradicts both the data from the world bank, which also opposes the data provided to us. We assume the data for this indicator has been altered at least twice and is therefore untrustworthy.

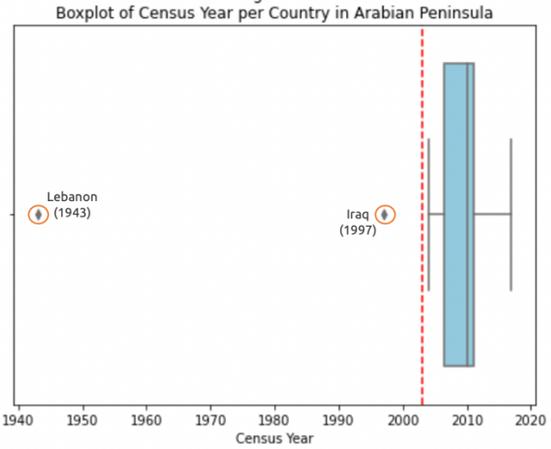
→ *Decision: Drop*

## 3.3 Literacy Rate

**(73.33% missing values)**: The meta data states that literacy rate is difficult to measure, and its definition and methods of data collection differ across countries and so should be used cautiously. Estimating these rates requires census measurements under controlled conditions. As such, we looked into the dates of the latest population census per country in Arabian Peninsula (from meta data) and found that:

- Years vary to a high degree
- Two outliers exist, Lebanon and Yemen with census reported in 1943 and 1997 respectively
- The average year of census reports for our region is 2010 (10 years old) (figure 8).

Figure 8:
Boxplot of Census Year per Country in Arabian Peninsula




→ *Decision: Drop*

These findings render the data untrustworthy.

# 4. Employment

**(No missing values)**: The seven indicators for the employment category shine by having a 100% fill rate, so no missing values. In addition, the indicators share the same trustworthy source, the International Labor Organization (ILO). However, there are strong factors in the metadata indicating that the data is not usable for an international comparison as conducted in this report:

- ILO models the indicators with data drawn from labor force surveys and supplements it with estimates, resulting in different reporting standards regarding definitions, coverage, and timelines
- Metadata states explicitly that these indicators have gender biases. Depending on different demographic, social, legal, and cultural trends and norms, it is differently determined whether women's activities are regarded as economic.
- 2/3 of the countries in our region are within the last 25 ranks of the latest Global Gender Gap Report (World Economic Forum, 2020) (50% within the last 15 ranks).

→ *Decision: Drop*

# 5 Environment

## 5.1 Energy Usage per GDP & GDP per Energy

**(13.33% missing values)**: Firstly, both data sets date back to 2014. Secondly, although the world bank displays the data for Energy Use and GDP per Energy, the link to their source is broken. After further research within the original source's database, we notice that the specific indicators cannot be identified due to different labels and a lack of metadata on our end.

→ *Decision: Drop*

## 5.2 Improved Water & Sanitation

**(No missing values)**: In 2015, the Joint Monitoring Programme by WHO/UNICEF segmented these two indicators into:

- People using safely managed sanitation services (% of the population) (SH.STA.SMSS.ZS)
- People using essential sanitation services (% of the population) (SH.STA.BASS.ZS)
- People using safely managed drinking water services (% of the population) (SH.H2O.SMDW.ZS)
- People using essential drinking water services (% of population) (SH.H2O.BASW.ZS).

Therefore, our data is outdated by five years, and since we're not allowed to import new columns for our analysis, we cannot conduct an analysis of the new indicators.

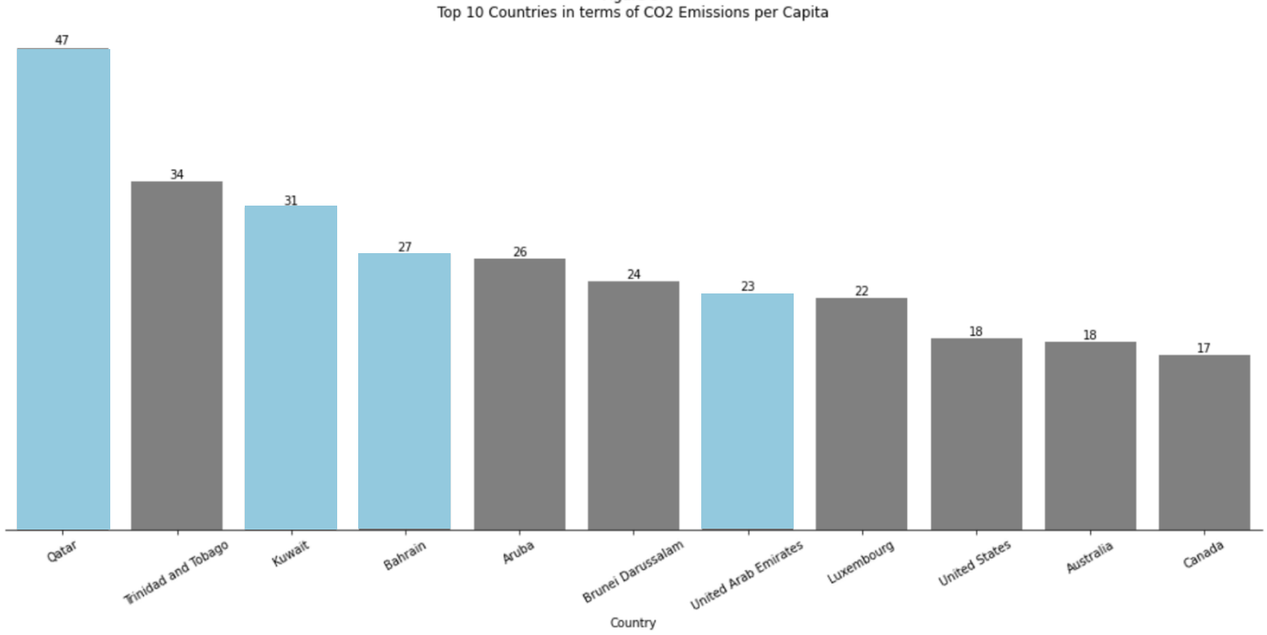
→ *Decision: Drop*

## 5.3 CO2 Emissions

**(No missing values)**: Although the data goes up to 2014, it is accurate and verifiable at its source. However, it's important to note a discrepancy of almost x3 between the numbers from the source and the numbers we have. This might be because the original data is expressed as metric tons of carbon, whereas our data isn't specified whether it's metric tons of carbon or carbon dioxide. Regardless of the discrepancy, the trend and ranking of countries from both data sets are similar: Qatar, Kuwait, Bahrain, and UAE are amongst the top 10 countries in the world with the highest CO2 Emissions per Capita (figure 9) (Asmakh & Al-Awainati, 2018).

→ *Decision: Keep*

Figure 9:
Top 10 Countries in terms of CO2 Emissions per Capita



# 6. Health

According to (Cammett, et al., 2014), the World Health Organization relies on government reported information. Due to the unstable governmental systems of most countries in the Arabian Peninsula, health organizations provide estimates or perceptions of the populations.

From 1980 to 2003, the countries in the regions have been implementing a system for registering vital events such as deaths and births. Depending on when the system was implemented, missing data has been imputed with estimated ages (U.S. Department of Health & Human Services, 2015). Due to the lack of access to appropriate health care, many cases are not appropriately reported.

Religion also plays an essential role in the lack of sex education and safety (Dupont, 2017)

## 6.1 Aids Deaths, ART Coverage, HIV cases

**(66.67% missing values)**: A proportion of HIV cases would results in deaths due to Aids due to no treatment (or late discovery). Art coverage indicates the percentage of all people living with HIV who are receiving therapy (HIV.gov, 2020).

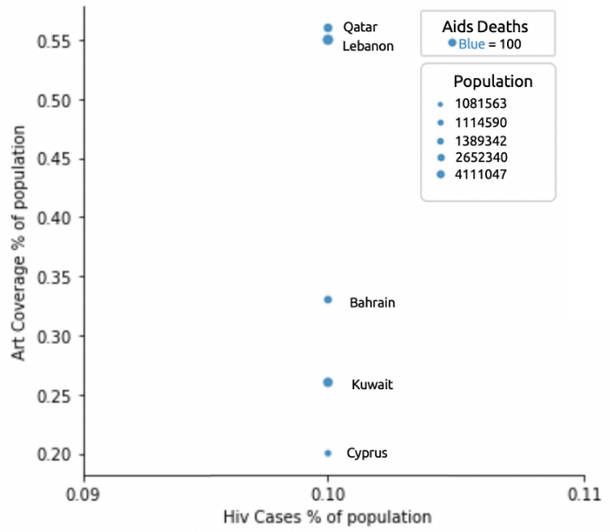
The numbers in the sheets are estimates by UNAIDS. Nothing is known about the collection of data.

If more people are aware of the problems through sex education (which is limited) or by more receiving more treatment, the cases and deaths should decrease in relation to the total population (Avert, 2020). There is no correlation with aids deaths, HIV cases, and treatment (as there should), so the estimates are not an accurate representation (GBD 2015 HIV Collaborators, 2016). This is also shown in figure 10. There is no linear correlation.

Lastly, due to the relative low number of people that get treated, it is also hard to indicate whether this the death was from aids or HIV.

→ *Decision: Drop*

Figure 10:
Correlation HIV Cases, Treatment, Aids Deaths




## 6.2 Malaria Cases

**(66.67% missing values)**: The data collection for the cases is unclear, however, the data does seem to be accurate in its proportion, looking at figure 11 (Zamani, et al., 2013).

The supplemented research study explains that the majority of the countries are not suitable for the growth of malaria. The missing values represent the countries that don't have suitable climates. Yemen in this case is an outlier, which is also shown in figure 13.

→ *Decision: Keep & impute with 0*

It is decided to impute the NaN values with 0, as the cases in the other countries are very unlikely and it would not make too much sense to impute them with the mean or median. As expected, the imputed distribution in figure 12 peaked much higher and the original distribution.

Figure: 11
Suitable Malaria Environments (Dark Grey)

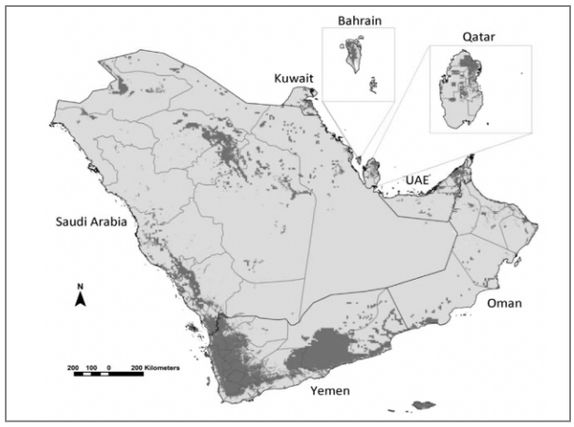



Figure 12:
Malaria Cases Distribution
(With/without imputation of missing values)

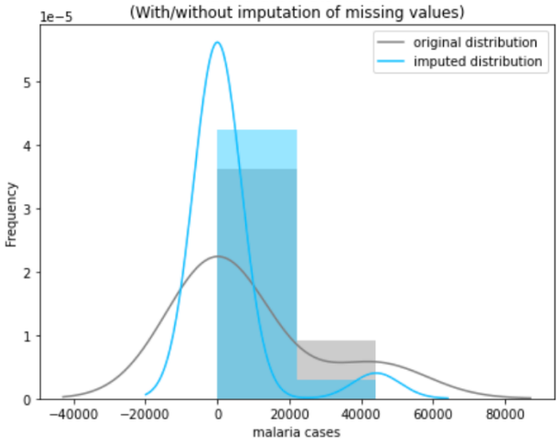
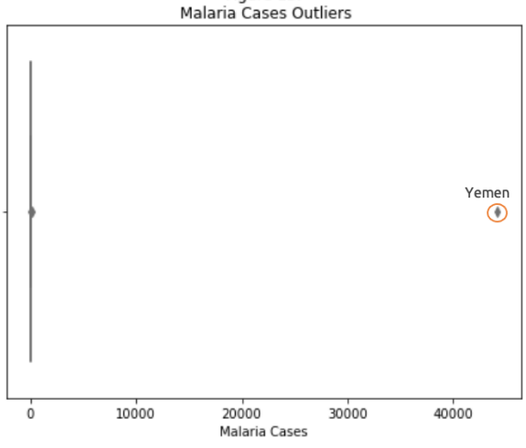



Figure 13:
Malaria Cases Outliers



## 6.3 Undernourishment

**(26.67% missing values)**: The indicator represents the percentage of the population whose food intake is insufficient to meet dietary energy. The Accuracy of the indicator is questionable and has its limitation according to the metadata:

1. Food insecurity is not a problem of inadequate access to food (regardless of availability)
2. Average food available to each person, even corrected for possible effects of low income, is not a good predictor of food insecurity.
3. Nutrition security depends on the quality of care of mothers and children and the household's health environment.

As the information on access to food, the average availability of food per person, and the quality of care is not in our data available as a separate indicator; it is hard to interpret the actual meaning of the percentage.

→ Decision: Drop

## 6.4 Tuberculosis Cases & Mortality

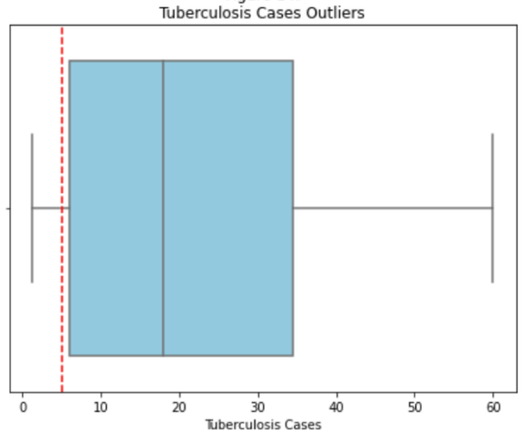
The report mentioned in the date meta states that new information on Tuberculosis (in combination with HIV) becomes more available every day. The report and country outlook on the original source shows that the numbers (in form of the weighted average, matches with the data reported in our data set. It also shows that most countries in our region has received support of treatment within the last five years (WHO, 2020).

*→ Decision: Keep both*

### 6.4.1 Tuberculosis Cases (No missing values)

The cases is an the estimated number of cases expressed as the rate per 100,000 population. Figure 14 shows there are no obscure outliers, however, there are two: United Arab Emirates (1.6) and Westbank and Gaza (1.3). It is important to keep those two in mind when comparing it with the other countries, therefore a threshold of 5.0 has been set. These countries are supposed to have better supportive systems for Tuberculosis (WHO, 2020).

Figure 14:
Tuberculosis Cases Outliers




### 6.4.2 Tuberculosis Mortality (6.67% missing values)

The meta data describes that the Tuberculosis death rate exist of tuberculosis among HIV-negative people, expressed as the rate per 100,000 population. As only one value is missing, it is best to impute it with the median looking at figure 15 and 16.

When looking at the outliers in figure 17, we can see that Yemen has higher rate for Tuberculosis Mortality. According to (International Association for Medical Assistance to travelers, 2020), Yemen is currently experiencing a high endemic of Tuberculosis, which explains why this country is considered to be an outliers.

Figure 15:
Distribution of Tuberculosis Mortality

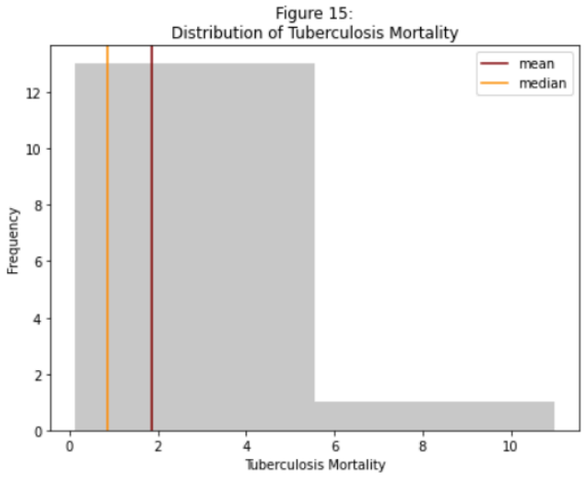



Figure 16:
Tuberculosis Mortality Distribution
(With/without imputation of missing values)

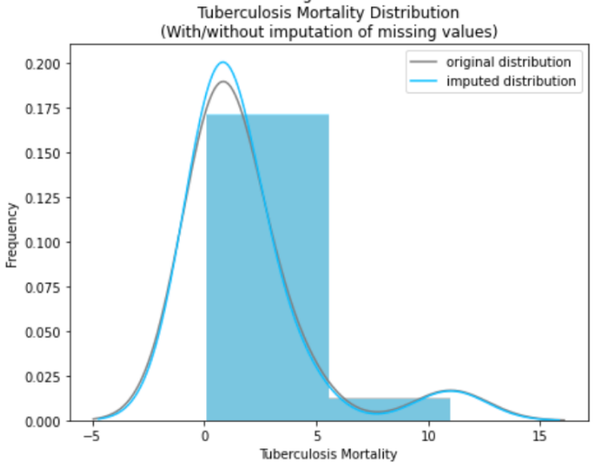
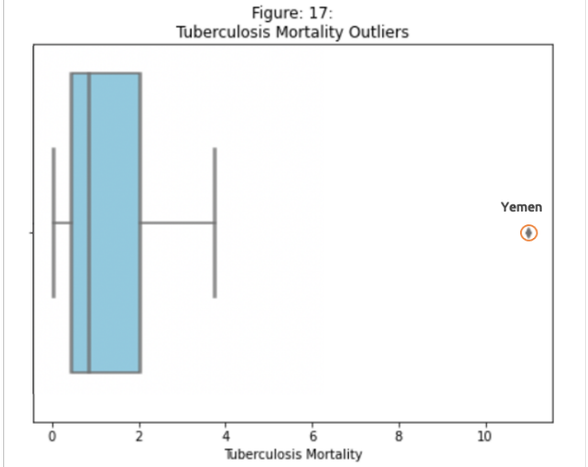



Figure: 17:
Tuberculosis Mortality Outliers



## 6.5 Measles Immunization

**(6.67% missing values)**: The meta data describes that the data from WHO was gathered from national censuses and nationally representative household surveys. *The last census data collection year varies within our region from 1943 to 2017.* In this case, the outdated censuses do not reflect an accurate representation of the immunization situation. However, more recent data has been published on (WHO, n.d), showing that the reported number in our data set does reflect more recent reports from 2019. This means that the data does still reflect the situation, regardless of the census collection year.

→ *Decision: keep*

With only one missing value, looking at figure 18 and 19, the *mean* would better represent the missing value. *Two outliers were found in figure 20, which are significantly lower than the rest: Yemen and Iraq.* As the box plot did not mark

Figure 18:
Distribution of Measles Immunization

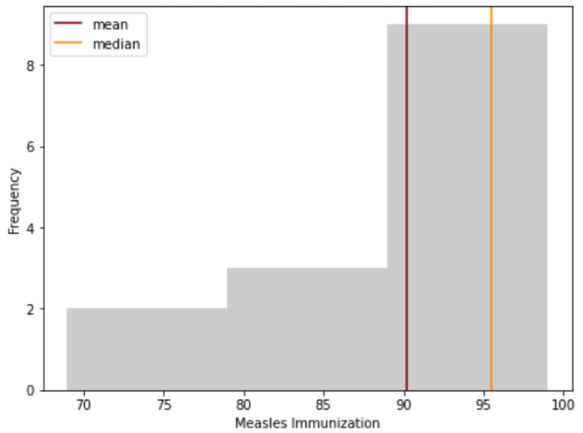



Figure 19:
Measles Immunization Distribution
(With/without imputation of missing values)

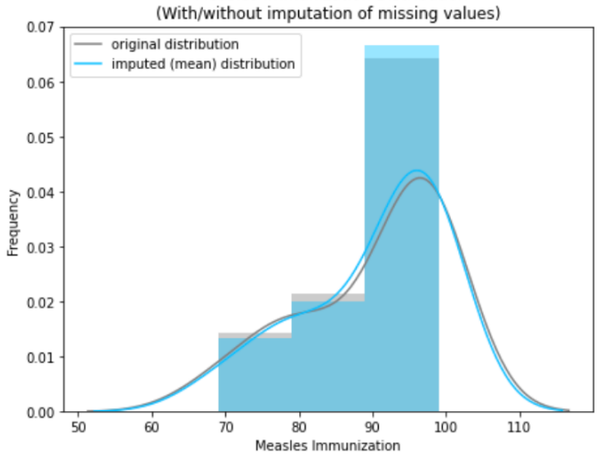
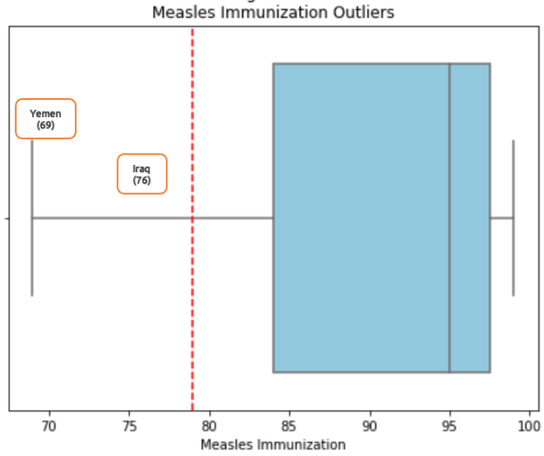



Figure 20:
Measles Immunization Outliers




## 6.6 Life Expectancy & Fertility Rate

**(No missing values)**: According to the metadata, the annual data series data are interpolated data from 5-year period data. The data was taken from six different sources to complement each other:

1. United Nations Population Division. World Population Prospects: 2019 Revision.
2. Census reports and other statistical publications from national statistical offices.
3. Eurostat: Demographic Statistics.
4. United Nations Statistical Division. Population and Vital Statistics Reports (various years).
5. U.S. Census Bureau: International Database.
6. Secretariat of the Pacific Community: Statistics and Demography Programme.

In general, it is more difficult to measure the data's accuracy due to the instability in some of the countries. To indicate the life expectancy and fertility rate, it is needed to combine different sources, given the situation.

→ Decision: Keep

There are two outliers in figure 21, Yemen and Iraq, which have lower life expectancy age. There are no outliers for the fertility rate (figure 22).

Figure 21:
Life Expectancy Outliers

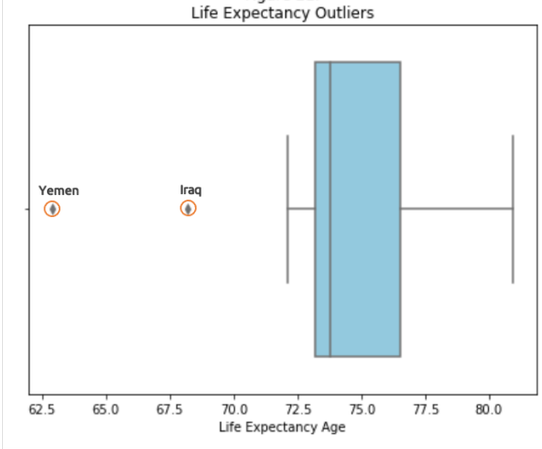
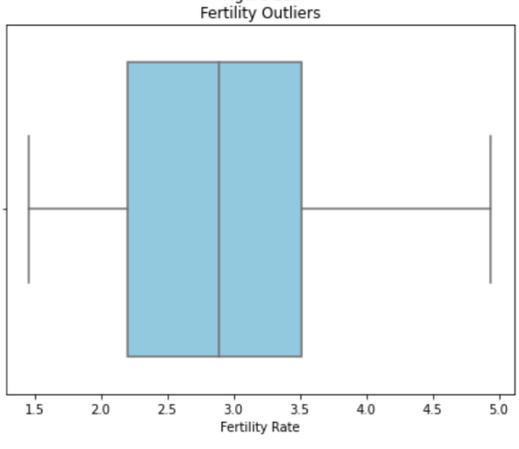



Figure 22:
Fertility Outliers



## 6.7 Maternal Mortality

**(No missing values)**: The metadata described that the ratios are generally of unknown reliability, and therefore it cannot be assumed that the provided ratios represent accurate estimates. There are other dependencies for mortality, which makes it difficult to measure. Therefore, our team cannot trust the data's accuracy and has decided to drop this indicator.

→*Decision: Drop*

## 6.8 Prenatal Care, Delivery Care, and Infant Mortality

Prenatal Care (80% missing values)
Delivery care (46.67% missing values)
Infant Mortality (no missing values)

Prenatal care refers to the percentage of pregnant women attended by skilled workers at least once during pregnancy. Delivery Care, on the other hand, refers to births attended by skills staff as a percentage. As described before and plotted in 3.3 Literacy Rate, the data of these indicators are dependent on censuses data.
As there are no other data substitutes, justifying the numbers and direct reasons for missing values is difficult.

→ *Decision: Drop*

## 6.9 Adolescent Fertility

**(No missing values)**: The source mentioned in the metadata does not has explicitly data related to Adolescent Fertility. There are data excel sheets with specific numbers per age; however, the average-weighted data does not match the data given in our dataset. Furthermore, there are no other sources or info on the collected data and cannot verify the accuracy.

→ *Decision: Drop*