Project Dataset Selection
Yemen : Cholera Outbreak Epidemiology Data

About Data

Yemen's cholera episode - the most exceedingly awful event on the planet - is once more in news , with approximately 10,000 speculated cases currently announced every week, the most recent information from the World Health Organization (WHO) appeared on recently.

Since Yemen's cholera pestilence ejected in April 2017, a sum of 1.2 million speculated cases have been accounted for with 2,515 deaths. Kids represent 30 percent of diseases.

Who (company, agency, organization) collected the data?

The dataset is provided by the Governorate of Yemen with the help of WHO(World Health Organization) over the years after the Cholera outbreak in 2017.

Who they are, what do they do? What is their role/purpose?

This dataset provides numbers of suspected cholera cases, deaths, case fatality rates (%) and attack rates (per 1,000) by Governorate in the Yemen cholera outbreak since 27 April 2017.

The data are manually extracted from the Yemen cholera outbreak epidemiology updates produced by WHO Yemen. The updates are posted on the Yemen Situation Reports page on the WHO Regional Office for the Eastern Mediterranean site. This dataset contains data from the daily and weekly updates up to the weekly report published on 7 July 2018.

This dataset contains figures from EPI bulletins, the one where the weekly EPI bulletins and daily bulletins are updated.

- Starting 19 June 2017, the unit used for the attack rate changed from per 10,000 to per 1,000. Previous data were adjusted to the new unit by dividing by 10.
- Starting 2 July 2017, the data included figures for a Moklla. These data were not mapped into any of the governorates in the Yemen CODs.
- Starting 6 July, the data included all the new figures. These data were not mapped into any of the governorates in the Yemen CODs.
- Governorate level updates were discontinued in this dataset. The last governorate level update was on 18 February 2018. The governorate level updates were discontinued in this dataset because the governorate level data is published as an image since that date.

Why did they collect this data?

The WHO with Yemen Governorate collected this data to keep a check on the growing cases of cholera and to arrange medical aid for the ones affected by the disease and to prevent the other from getting the disease. The government wanted to make sure that the affected ones were given urgent and proper medical facilities at the earliest and to curb the oncoming effects of the disease as soon as possible. So, in order to do that they collected all the data related to the disease to deal with the situation with all arms.

Why is this a big data problem?

The problem is a big data problem as the impact of the disease was spread all over the country with thousands of cases increasing day by day. Thus, the inflow of data was in huge amounts with several variables to dig deep into.

Deaths, fatality rates, number of cases, fatality ratios and several other deliverables needed to be analyzed to deal with the situation. This constituted it to be a big data problem.

Describe any privacy, quality, ethical, or other issues with this dataset?

The issues that will be faced with the dataset is that it is not a clean dataset. It has several values missing in a couple of columns. The dataset needs to be clean before any kind of analysis is applied on it to get a proper outcome from the data. The missing data deteriorates the quality of analysis and outcomes and can result in inappropriate treatment which could lead to deterioration in progress of the case and can also lead to a higher death rate.

What potential questions could be answered by studying this data?

Several questions can be asked and answered by studying the data some of which can be :

1. How many number of cases of cholera were cited and how many deaths happened of those cited cases?
2. What was the rate of disease attacks that have occurred when compared to the attacks causing deaths?
3. How many governorate were the most affected ones by cases of cholera?
4. Which ones were the highest and lowest affected governorate as well?
5. How were the case fatality ratio and the cases that were cited related?
6. What were the Pcodes that were related to the greatest number of deaths.
7. What was the relationship between cases of disease cited and deaths and find was there any pattern?
8. What were the areas which held the greatest number of cases of deaths?
9. What dates witnessed the highest number of deaths?
10. How many deaths happened on a particular date?
11. Did the deaths increase with time?

And a lot more questions can be asked and answered…

Requirements:
Software: SQL, Python and R.
Hardware: Laptop with 8gb ram, SSD with at least 50gb storage space.

About the dataset:
The dataset has 5432 rows and 7 columns.
Size: 300kb
Name: yemen-governorate-level-cholera-epidemiology-data-csv-1.csv

Metadata: Description of Dataset Columns

Column	Description
Date	Date when the figures were reported.
Governorate	The Governorate name as reported in the WHO epidemiology bulletin.
Cases	Number of cases recorded in the governorate since 27 April 2017.
Deaths	Number of deaths recorded in the governorate since 27 April 2017.
CFR (%)	The case fatality rate in governorate since 27 April 2017.
Attack Rate (per 1000)	The attack rate per 1,000 of the population in the governorate since 27 April 2017.
COD Gov Pcode	The PCODE name for the governorate according to the Inter Agency Standing Committee (IASC) Common Operation Datasets (CODs) for Yemen.

Data Dictionary

S.No.	Column : Variables	Data Type
1.	Date	Date
2.	Governorate	String
3.	Cases	Integer
4.	Deaths	Integer
5.	CFR (Case Fatality Ratio)	Decimal
6.	Attack Rate (Per 1000)	Decimal
7.	COD Gov Pcode	Integer

Citation and References:

[1] Yemen : Cholera Outbreak Epidemiology Dataset – Retrieved from - https://data.world/hdx/c12b27ea-5be4-4064-b87c-624cbbb7b1e1