ETL Project

Title: Prevalence of mortality of the three international outbreaks.

Why?

First, we looked for a relevant topic in recent times, and, which topic is more relevant than diseases?

Extraction:

Extraction of Data from: https://www.who.int/data/gho, https://www.who.int/emergencies/diseases/novel-coronavirus-2019

We extracted the data as csv files and we stored it on a directory called “Resources” as covid_2020.csv, h1n1_2009.csv and sars_2003.csv.

Transformation:

The transformation of data: We extracted the data from the csv files and created data frames using pandas “pd. read_csv(file.csv)”. We selected the columns needed for the analysis such as country, cases, deaths. Whenever it was needed, we renamed the columns to match the afore mentioned variables. Then we selected only the countries with cases >0. The we group them by country, so we could give the data frames the format to merge it, using as a primary key = “country”. Afterwards we calculate the number of cases worldwide, the same for the deaths, and then we where able to calculate the death ratio which is the (# death*100/# cases).

Load:

We used choropleth to graph the prevalence and mortality worldwide. We saved it as a csv. And we called it c_h_s.csv, and that’s our final database.

Even though only two or three members of the team will send the url for grading the activity, all the following members work together to get this done:

- Daniely Miranda
- Enrique Gaspar
- Erick Castillo
- Federico Mendoza Renaud
- Hector Contreras Secchi