**Peru Covid-19 Data Web Scrape**

Compiled by Boyer Simpkins on 09/08/2022

All the data collected via this web scrape is found at Peru’s National Platform of Open Data, (<https://www.datosabiertos.gob.pe/search/field_topic/covid-19-917?sort_by=changed>). The data sets of interest are Covid-19 deaths, positive cases, and deaths, hospitalizations, and vaccines. There is a plan in place to include vaccination data though access is currently denied to that data set. All the data was collected using python and the request packages. I will go into more detail about how I cleaned the data and the logic behind it for each data set. Our current plan is to collect these three data sets before dumping raw and cleaned data into a database. As it currently stands, this would entail the maintenance of six tables: two for each data set. Once we can upload the data, we plan on developing a batch file that will allow us to run the web scraping algorithm every morning. Below is the general description of the data.

**Deaths:** (<https://www.datosabiertos.gob.pe/dataset/fallecidos-por-covid-19-ministerio-de-salud-minsa>) Total deaths due to Covid 19 currently recorded throughout Peru.

*Size:*

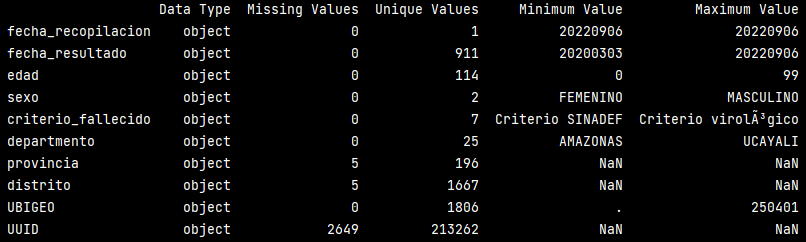
1. Raw data: 215,982 Rows x 10 Columns.
   1. 

Figure 1: Data Quality Report for Raw Deaths data.

1. Cleaned Data: 213,258 Rows x 10 columns.
   1. Text

      Description automatically generated

Figure 2: Data Quality Report for cleaned data.

* 1. In order to clean the data, I converted fecha\_recopilacion and fecha\_resultado into datetime objects. I also removed the five NA values found in provincia and the 2,649 found in UUID. There may be a way to assign a previously unused UUID to these missing values, but as of right now I am just dropping them since these data points represent less than 1% of the total data. I also removed duplicate rows.

*Features:*

1. Fecha\_Recopilacion: date of data collection. This date will be the same for each row in the data set. Date format = DD/MM/YYYY. (MM/DD/YY)
2. Fecha\_Resultado: date of results. In the case of deaths, it is the date that individual died. Date format = DD/MM/YYYY, (MM/DD/YY)
3. Edad: age of the individual in years.
4. Sexo: sex of the individual. Only has a binary representation with values of masculino or feminino for male and female respectively.
5. Criterio\_Fallecido: the criteria of death which is defined by the CDC with the following values.
   1. Virological criterion: Death in a confirmed case of COVID-19 who dies within 60 days after a molecular (PCR) or reactive antigen test for SARS-CoV-2.
   2. Serological criterion: Death in a confirmed case of COVID-19 who dies within 60 days after a positive IgM or IgM/IgG serological test for SARS-CoV-2.
   3. Radiological criterion: Death in a probable case of COVID-19 that presents a radiological, tomographic or nuclear magnetic resonance image compatible with COVID-19 pneumonia.
   4. Epidemiological link criterion: Death in a probable case of COVID-19 that has an epidemiological link with a confirmed case of COVID-19.
   5. Epidemiological investigation criteria: Death in a suspected case of COVID-19 that is verified by epidemiological investigation of the National Epidemiology Network (RENACE).
   6. Clinical criteria: Death in a suspected case of COVID-19 with a clinical picture compatible with the disease.
   7. SINADEF Criteria Death with death certificate in which the diagnosis of COVID-19 is presented as the cause of death. Death due to COVID-19 on the death certificate is defined by the presence in fields A, B, C or D of the ICD-10 codes: U071, U072, B342, B972, or the mention of the terms "coronavirus" , “cov-2”, “cov2”, “covid” and “sars”.
6. Departamento: the department or region where the individual was located. It is analogous with a state feature when comparing against the USA. There are twenty-five unique departments.
7. Provincia: the province where the individual was located. It is analogous with a county feature when comparing against the USA. There are 196 unique provinces with each one belonging to a department except for the Lima Province.
8. Distrito: the district where the individual was located. It is analogous with a city feature when comparing against the USA. There are 1,838 unique districts ranging in population from 1.1 million to 3,500.
9. UBIGEO: Geographic location code denoting “DDppdd” for (department, province, and district respectively). Source for these codes in INEI (Instituto Nacional de Estadística e Informática) which is a government organization in Peru tasked with evaluating statistical information in the country.
10. UUID: a unique ID for each individual.

**Positive Cases:**

*Size:*

1. *Raw data:* 4,126,021 Rows x 10 Columns

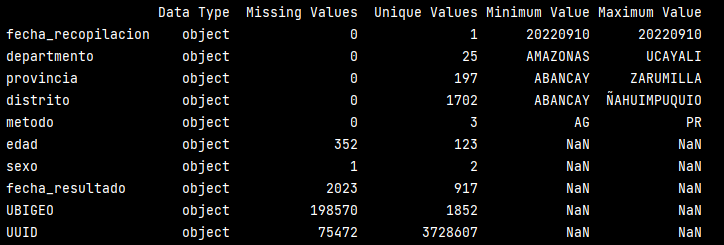


Figure 3: Data Quality Report for Raw Positive Cases.

1. Cleaned Data: 3,770,591 Rows x 10 Columns

Graphical user interface, text

Description automatically generated

Figure 4: Data Quality Report for Cleaned Positive Cases.

As of now I am currently just dropping all NA values since they should all have some type of value. That being said, the first times I was running this script there was a lot less NA values then is currently being reports, so some type of additional analysis may need to be done. I also dropped all duplicate rows.

*Features:*

1. UUID: Same feature description as used in deaths.
2. Fecha\_Recopilacion: Same feature description as used in deaths.
3. Fecha\_Resultado: Same feature description as used in deaths.
4. Edad: Same feature description as used in deaths.
5. Sexo: Same feature description as used in deaths.
6. UBIGEO: Same feature description as used in deaths.
7. Departamento: Same feature description as used in deaths.
8. Provincia: Same feature description as used in deaths.
9. Distrito: Same feature description as used in deaths.
10. Metodo: The method of test that was used in order to confirm the positive case.

**Deaths, Hospitalizations, and Vaccines** (DHV, <https://www.datosabiertos.gob.pe/dataset/fallecidos-hospitalizados-y-vacunados-por-covid-19>): The cross section between deaths, vaccination status, and hospitalization status throughout Peru.

*Size:*

1. Raw Data: 119,980 Rows x 35 Columns.
   1. Text

      Description automatically generated

Figure 5: Data Quality Report for Raw DHV data.

1. Cleaned Data: 119,972 Rows x 31 Columns:
   1. Text

      Description automatically generated with low confidence

Figure 6: Data quality Report for Cleaned DHV Data.

* 1. There is a lot more cleaning that could be done, and I am awaiting on more instruction before I proceed. That being said, I filtered out the last four rows from the raw data as they were redundant features. As you can see there are still a lot of NaN values found in the data, but they were not removed because they hold some significance. For instance, fecha\_dosis1 is the date that an individual was first vaccinated. If the individual was not vaccinated than it is represented with a NaN values. I will go into more detail about this in the feature description. I also removed duplicate rows.

*Features:*

1. UUID: Same feature description as used in deaths and positive cases.
2. Fecha\_Recopilacion: Same feature description as used in deaths and positive cases.
3. Fecha\_Resultado: Same feature description as used in deaths and positive cases.
4. Edad: Same feature description as used in deaths and positive cases.
5. Sexo: Same feature description as used in deaths and positive cases.
6. Criterio\_Fallecido: Same feature description as used in deaths and positive cases.
7. UBIGEO: Same feature description as used in deaths and positive cases.
8. Departamento: Same feature description as used in deaths and positive cases.
9. Provincia: Same feature description as used in deaths and positive cases.
10. Distrito: Same feature description as used in deaths and positive cases.
11. CDC\_Positividad: If a case was confirmed as positive or not by the CDC. A value of zero represents that a positive result was not confirmed by the CDC while a value of one represents that a positive result was confirmed by the CDC.
12. Flag\_Vacuna: Represents the number of doses an individual has received of a vaccine. The lowest value being zero and the highest current value being three.
13. Fecha\_Dosis1: The date that the first dose of a vaccine was administered. A value of NaN corresponds an individual who never received a vaccination.
14. Fabricante\_Dosis1: The is the manufacturer of the first dose that the individual received. A value of NaN corresponds an individual who never received a vaccination.
15. Fecha\_Dosis2: The date that the second dose of a vaccine was administered. A value of NaN corresponds an individual who never received a vaccination or a second dose. The individual may have received a dose, however.
16. Fabricante\_Dosis2: The is the manufacturer of the second dose that the individual received. A value of NaN corresponds an individual who never received a vaccination or a second dose. The individual may have received a dose, however.
17. Fecha\_Dosis3: The date that the third dose of a vaccine was administered. A value of NaN corresponds an individual who never received a vaccination or a second dose or a third dose. The individual may have received two doses, however. ( not clear)
18. Fabricante\_Dosis3: The is the manufacturer of the third dose that the individual received. A value of NaN corresponds an individual who never received a vaccination or a second dose or a third dose. The individual may have received two doses, however.
19. Flag\_hospitalizado: A binary representation of if an individual was hospitalized or not. A value of zero corresponds to no hospitalization while a value of one corresponds to a hospitalization.
20. Eess\_renaes: Codes of the establishment where the individual was hospitalized. A value of zero corresponds to someone who was not hospitalized.
21. Eess\_diresa: Codes of the establishment where the individual was hospitalized. NaN value corresponds to someone who was not hospitalized.
22. Eess\_red: Codes of the establishment where the individual was hospitalized. NaN value corresponds to someone who was not hospitalized.
23. Eess\_nombre: Codes of the establishment where the individual was hospitalized. NaN value corresponds to someone who was not hospitalized.
24. Fecha\_ingreso\_hosp: The date that the individual was hospitalized. NaN value corresponds to someone who was not hospitalized.
25. Flag\_UCI: Denotes if an individual was admitted to intensive or neonatal intensive care. A value of zero refers to someone who was admitted to the ICU or NICU while a value of one refers to someone who was not admitted to the ICU or NICU.
26. Fecha\_ingreso\_uci: The date when the individual was admitted to the ICU. A NaN value refers to an individual who was never admitted to the ICU.
27. Fecha\_ingreso\_ucin: The date when the individual was admitted to the NICU. A NaN value refers to an individual who was never admitted to the NICU. There may be some inconsistencies as there are many elderly individuals who are listed as being admitted to the NICU which I interpret as only being for infants.
28. Con\_Oxigeno: If a hospitalization required oxygen or not, with a value of zero being no oxygen required while a value of one being oxygen required.
29. Con\_Ventilacion: If a hospitalization required a ventilator or not, with a value of zero being no ventilator required while a value of one being oxygen required.
30. Fecha\_Segumiento\_hosp\_ultimo: Last date of monitoring at the hospital. A value of NaN corresponds to an individual who was never admitted to the hospital.
31. Evolucion\_hosp\_ultimo: Status of the individual on the last date of monitoring at the hospital. NaN values should refer to people who were not admitted to the hospital. Though there is a small subset of people, around one hundred, who were admitted but there was no status listed on the last day of their hospital visit. There are seven unique statuses,
    1. Alta\_voluntaria: Voluntary discharge.
    2. Defunción: The individual died in the hospital.
    3. Desfavorable: Discharged from the hospital with an unfavorable evaluation.
    4. Estacionario: Discharged in stable condition.
    5. Favorable: Discharged in favorable condition.
    6. Alta: Medical discharge.
    7. Referido: Referred to another medical center.
    8. Vacío: No medical record of the individual, which did not appear in this data set.