Presented to the Department of Software Technology

De La Salle University - Manila

Term 2, A.Y. 2021-2022

In partial fulfillment

of the course

Data Preparation and Integration - DATAPRE

COVID-19 Data Drop - Data Dictionary

Submitted by:

Florin, Adrian D.

Santiago, Jairus Ramiel C.

Submitted to:

Ms. Unisse Chua

June 17, 2022

Introduction

This document contains the data collected from the Philippine Department of Health's (DOH) COVID-19 Data Drop, which was used for the performed exploratory data analysis. The DOH COVID-19 Data Drop includes several other datasets pertaining to the data gathered by the DOH during the spread of the pandemic across the Philippines. The data drop includes weekly reports, case information, and quarantine facility data, which is collected by the DOH among respective health facilities and testing laboratories.

The data has varied types, ranging from floats to dates, which serve as the patient's personal information and COVID-19 status. The goal of this project is to organize and explore data in an effort to understand and implement the data science and preparation concepts discussed this term in the form of a publication-ready dataset derived from the previously mentioned COVID-19 data drops produced by the Philippines' Department of Health. This document's purpose is to introduce and expand on the data columns to be used in the final dataset of the columns found in the source. Furthermore, the data drops are freely available in the DOH's COVID Tracker for public use (https://doh.gov.ph/covid19tracker).

As of June 17, 2022, the current dataset has the following features that are listed below. It should be noted that any further changes to the existing variables in the dataset for its final publication shall be reflected in this document.

Data Columns

A. Patient Information

Patient Case Code
(*CaseCode)*

a. Data Type: Text/string
b. Description:

This field contains the randomized code provided to the patient to label the case. The identified code does not necessarily pertain to the unique case number assigned by the DOH to uphold the anonymity of the patient. Due to the nature of how the code was randomly generated, some case codes may be duplicated in the dataset. Nonetheless, each case provided in the dataset is unique in terms of features that are attributed to the case.

Patient Age
(Age)

a. Data Type: Number/float64
b. Description:

This field contains the age of the patient on the date when the testing specimen was collected. These ages are then grouped according to five-year intervals in the next data column '*AgeGroups*'.

Patient Age Groups
(AgeGroups)

a. Data Type : Text/String
b. Description

This field is used to classify cases according to age groups, with each group encompassing five years (formatted as 15 to 19, 25 to 29, etc.) with the lowest recorded age being 10 and the highest group being labeled "80+".

Sex
(Sex)

a. Data Type: Text/String
b. Description:

This field is used to classify patients according to their biological sex, between Male and Female.

Patient Pregnancy
(*Pregnanttab)*
Categorical Feature

a. Data Type: Text/String
b. Description:

This field indicates whether the patient is pregnant, with the applicable labels/data being “Yes”, “No”, or “N/A”. The data is also affected by the column ‘*Sex*’ that denotes the patient’s sex as if the person is Male, the observation is automatically filled with an “N/A”.

B. COVID-19 Case Registration and Status

Date of Onset Symptoms
(DateOnset)

a. Data Type: datetime64
b. Date format: yyyy-mm-dd
c. Description:

In this column, the data describes the date of when the onset symptoms of COVID-19 started to appear. Onset symptoms are defined as the first appearance of signs or symptoms of an illness. This data was presumably taken to provide a starting date of when the virus has infected a person.

Date of Specimen Collected
(DateSpecimen)

a. Data Type: datetime64
b. Date format: yyyy-mm-dd
c. Description:

The data in this column is the date of which each COVID-19 specimen is collected from a particular patient.

Date of Result Release
(DateResultRelease)

a. Data Type: datetime64
b. Date format: yyyy-mm-dd
c. Description:

This field contains the date at which the results of the test for the collected specimen was released.

Date Publicly Announced as a Confirmed Case
(DateRepConf)

a. Data Type: datetime64
b. Date format: yyyy-mm-dd
c. Description:

This field contains datetime data that indicates the date the patient has been confirmed and declared as another COVID-19 case and added to the list of confirmed cases.

Date of Death
(DateDied)

a. Data Type: datetime64
b. Date format: yyyy-mm-dd
c. Description:

This field contains the date at which the patient has died. The field is empty if the patient did not succumb whilst considered as an active case.

Date Recovered from COVID-19
(DateRecover)

a. Data Type: datetime64
b. Date format: yyyy-mm-dd
c. Description

This field contains the date at which the patient has been considered as recovered after testing negative for COVID-19.

Patient Health Status
(HealthStatus)
Categorical Feature

a. Data Type: Text/String
b. Description:

This column contains data that describes the known current status of the patient, with there being six different labels that can be used. These labels are as follows in order of severity, as lifted from the DOH's COVID-19 disease severity classification and corresponding intermediate package:

- Recovered - Tested negative for the virus after being admitted for quarantine.
- Asymptomatic - Tested positive for the virus but does not exhibit any signs and symptoms of the virus, e.g., cough, fever, etc.
- Mild - Patient is exhibiting symptoms such as cough, fever, anorexia, and myalgias. Patient also does not show any signs of pneumonia.
- Moderate - Patient shows signs of non-severe pneumonia, accompanied by cough, fever, and dyspnea (shortness of breath).
- Severe - Patient is experiencing severe pneumonia or severe acute respiratory infection accompanied by cough, fever, and dyspnea.

- Critical - Patient's condition is worsening a week after the initial test result. Signs of organ dysfunction/failure are also manifested through increased heart rate, low blood pressure, low oxygen saturation, and altered mental status.
- Died - Patient has expired.

Furthermore, this column's data affects the presence of data in other columns, particularly in the *DateDied* and *DateRecovered* columns.

C. Quarantine Status

Removal Type
(RemovalType)
Categorical Feature

a. Data Type: text/object
b. Description:

This field contains how the patient was relieved from the quarantine site if he/she has been quarantined by the local health center. The removal type is categorical and is labeled as recovered or died. The field is left empty if the case was not admitted into the local health facility for quarantine.

Admitted
(Admitted)
Categorical Feature

a. Data Type
b. Description:

This field indicates if the patient was admitted into a quarantine facility by the local health center after testing positive for the COVID-19.

Patient Home Quarantine Status
(*Quarantined)*
Categorical Feature

a. Data Type: Text/String
b. Description:

This column contains data that describes the patient's quarantine status. Put simply, the data indicates whether the patient has been home quarantined or not and is marked by either a "Yes" or "No".

D. Location Data

Region of Residence
(RegionRes)

a. Data Type: Text/String
b. Description

This field contains the region of residence of the patient, with the possible options being from the Philippines' 17 regions. It is formatted according to region numbers, an example being "Region IV-A: CALABARZON". There are also exceptions, and these are the regions that are labeled without the "Region" label. These include the NCR, CAR, and CARAGA regions.

Province of Residence
(ProvRes)

a. Data Type: Text/String
b. Description:

This field contains the province of residence of the patient, with the possible data being the 81 regions in the Philippines across all 17 regions. This field is empty if a region is not provided.

City of Residence
(CityMunRes)

a. Data Type: Text/String
b. Description:

This field contains the city of residence of the patient. This field is empty if a region or province is not provided.

Philippine Standard Geographic Code of Municipality for Cities
(CityMunPSGC)

a. Data Type: Text/String
b. Description:

This field contains the PSGC of the patient's city of residence. The PSGC is a coding system that classifies the geographical locations of the Philippines using a 4-layered hierarchy which, from highest to lowest, is classified as follows:

- Region
- Province
- City/Municipality
- Barangay/Village

The coding scheme of the PSGC uses 9 digits, the first two being the region number, the second pair being the provincial code, the third pair as the

municipality/city code, and the last three digits is the barangay code. This field is empty if the fields containing the Region, Province, or City of residence are also empty, as there is no location to assign a PSGC to.

Barangay of Residence
(BarangayRes)

a. Data Type: Text/String
b. Description:

This field contains the barangay of residence of the patient. This field is empty if a region, province, or city is not provided.

Philippine Standard Geographic Code of Municipality for Barangays
(BarangayPSGC)

a. Data Type: Text/String
b. Description:

This field contains the PSGC of the patient's city of residence. The PSGC is a coding system that classifies the geographical locations of the Philippines using a 4-layered hierarchy which, from highest to lowest, is classified as follows:

- Region
- Province
- City/Municipality
- Barangay/Village

The coding scheme of the PSGC uses 9 digits, the first two being the region number, the second pair being the provincial code, the third pair as the municipality/city code, and the last three digits is the barangay code. This field is empty if the fields containing the Region, Province, City, or Barangay of residence are also empty, as there is no location to assign a PSGC to.