Prediction of COVID-19 around the world

Student: Angela Amador

TMU Student Number: 500259095

Supervisor: Tamer Abdou, PhD

Submission Date: Oct 16th, 2023

**Table of Contents**

[**Abstract 2**](#_heading=h.30j0zll)

[**Literature Review 3**](#_heading=h.3znysh7)

[Machine Learning‑Based Research for COVID‑19 Detection, Diagnosis, and Prediction: A Survey [4] 3](#_heading=h.sjlzr9ijv6pu)

[Kalman filter based short term prediction model for COVID-19 spread [13] 5](#_heading=h.usjlvgxlmtb)

[Predicting the Growth and Trend of COVID-19 Pandemic using Machine Learning and Cloud Computing [8] 6](#_heading=h.4d34og8)

[Machine learning-based prediction of COVID-19 diagnosis based on symptoms [3] 6](#_heading=h.2s8eyo1)

[Forecast and prediction of COVID-19 using machine learning [6] 6](#_heading=h.26in1rg)

[Forecasting COVID-19 spreading through an ensemble of classical and machine learning models: Spain’s case study [9] 7](#_heading=)

[Conclusion 7](#_heading=h.qf2m5qgq1x5y)

[**Exploration Data Analysis 7**](#_heading=h.3vkygbp5fn5s)

[Dataset 7](#_heading=h.w88kulit1yoo)

[Data Dictionary 7](#_heading=h.ot5fgaf92fz2)

[Metadata 9](#_heading=h.3b23vrqpjytm)

[Dimensionality Reduction (CMTH642 - Module 9) 14](#_heading=h.jc36mxywvgh0)

[Removing data columns with too many NaN values 15](#_heading=h.n5fxmfjqkz63)

[Low Variance Filter 15](#_heading=h.igjg8pg6tu3)

[High correlation with other data columns 15](#_heading=h.x0udzqlv1wo)

[Summary of the techniques 16](#_heading=h.4puqwnetujr7)

[Dataset subset 16](#_heading=h.j9nhshl4a376)

[Overview 18](#_heading=h.8c6hksj4fe29)

[Variables 19](#_heading=h.b6jj0dwbwdg9)

[GITHUB Repository: 27](#_heading=)

[**References 27**](#_heading=h.9jhbyp3p5z76)

# Abstract

The novel Coronavirus disease (COVID-10) was first reported in December, 2019 in Wuhan, Hubei Province, China. It created a calamitous situation throughout the world as cumulative incidents of COVID-19 were rapidly increasing day by day. In the absence of any medications, the only solution was to slow down the spread by exercising “social distancing” (hard lock-downs, restrictions on people mobility, limitations of the number of people in public places and the usage of protection gear (masks or gloves), among others) to block the chain of the spread of the virus. Here it is where Machine Learning models helped to forest where and when the disease is likely to spread, and notify those regions, governance and entities on their decision making.

# Literature Review

Several publications and studies were reviewed with emphasis being placed on predicting the number of cases around the world and how these machine learning models helped the governance and organizations to better prepare for the pandemic.

## Machine Learning‑Based Research for COVID‑19 Detection, Diagnosis, and Prediction: A Survey [4]

This paper reviews more than 160 Machine Learning (ML) - based approaches developed to help with the pandemic. They include the type of the addressed problem (detection, diagnosis, or detection). The scope of this project is prediction, based on the analysis of the paper, these are the methods and data types that have been used for prediction:

Some of the supervised learning models for prediction of COVID-19 cases:

| **Method Name** | **Data Type** |
| --- | --- |
| Support Vector Machine (SVM) with Decision Tree (DT) | X-ray image |
| Support Vector Machine (SVM) | Text |
| Least Square-SVM (LS-SVM) and Autoregressive Integrated Moving Average (ARIMA) | Time series |
| Linear regression model and Random Forest | CT images |
| Logistic regression model | CT images |
| XGBoost | Time series |
| Linear regression model with Support Vector Machine (SVM) Model and Artificial Neural Network (ANN) | Text |
| Linear regression and SEIR (Susceptible, Exposed, Infectious, Recovered) | Time series |
| Logistic Regression with Random Forest, Partial Least Squares Regression (PLSR), Elastic Net and Bagged Flexible Discriminant Analysis (BFDA) | Time series |
| Support Vector Regression (SVR), Stacking Ensemble Learning (SEL), Auto-Regression Integrated Moving Average (ARIMA), Cubist Regression (CUBIST), Random Forest (RF), Ridge Regression (RIDGE) | Time series |
| Support Vector Regression (SVR), Linear Regression and Polynomial Regression | Text |
| Linear regression models (Penalized Binomial Regression (PBR), Conditional Inference Trees (CIR), Generalised Linear (GL), and SVM with linear kernel) | CT Images and clinical data |
| PBRR (combination of Bayesian Ridge Regression (BRR) with n-degree Polynomial for forecasting) | Text |
| Fine-tuned Random Forest model with AdaBoost algorithm | Text |

Some of the Convolutional Neural Networks (CNN) approaches for prediction of COVID-19 cases:

| **Method Name** | **Data Type** |
| --- | --- |
| DenseNet-121 | CT images |

Some of the Recurrent Neural Networks (RNN) approaches for prediction of COVID-19 cases:

| **Method Name** | **Data Type** |
| --- | --- |
| LSTM with NLP | Text |
| LSTM | Text |
| LSTM | Time series |

Specialized CNN approaches for prediction:

| **Method Name** | **Data Type** |
| --- | --- |
| COVID–SDNet | X-ray images |

Other Machine Learning approaches for prediction of COVID-19 cases:

| **Method Name** | **Data Type** |
| --- | --- |
| Autoregressive Integrated Moving Average (ARIMA) model and Wavelet-based forecasting (WBF) model | Time series |
| MAchine learning and Cloud Computing | Time series |
| FbProphet technique and Logistic Model | Time series |
| Kalman Filter model | Text |

## Kalman filter based short term prediction model for COVID-19 spread [13]

This article analyzes various studies using the latest data of COVID-19 spread which includes demographic and environmental factors to be used into different Machine Learning Models like minimum temperature, maximum temperature, humidity, and rainfall in India.

Kalman filter has been used to forecast COVID19. Pearson correlation has been used to find the dependency of different features of the data. The importance of the feature in the proposed model has been calculated through the random forest algorithm.

The article concluded the proposed prediction model is good for short term prediction i.e. daily and weekly. The proposed prediction model can be updated to accommodate long term and medium term series prediction in future.

## Predicting the Growth and Trend of COVID-19 Pandemic using Machine Learning and Cloud Computing [8]

The focus of this article in addition to Machine Learning is Cloud Computing and how the power of Cloud Computing helped with the process to develop, manage and analyse big data. Cloud computing can be used to rapidly enhance the prediction process using high-speed computations.

The focus is to show that using iterative weighting for fitting Generalized Inverse Weibull (GIW) distribution, a better fit can be obtained to develop a prediction framework.

## Machine learning-based prediction of COVID-19 diagnosis based on symptoms [3]

This paper proposed a machine-learning model that predicts a positive SARS-CoV-2 infection in a RT-PCR test by asking eight basic questions. The model was trained on data of all individuals in Israel tested for SARS-CoV-2 during the first months of the COVID-19 pandemic. The model was implemented globally for effective screening and prioritization of testing for the virus in the general population.

Because the data is coming from surveys it has limitations, biases and missing information. Training and testing a model while filtering out symptoms of high bias in advance still achieved very high accuracy. The methodology presented in this study may benefit the health system response to future epidemic waves of this disease and of other respiratory viruses in general.

Predictions were generated using a gradient-boosting machine model built with decision-tree base-learners.

## Forecast and prediction of COVID-19 using machine learning [6]

The article discusses Auto Regressive Integrated Moving Average (ARIMA) time series for

forecasting confirmed cases for various states in India. Two classifiers, Random Forest

and Extra Tree Classifier (ETC), were selected. These results can be used to take corrective measures by different government bodies.The availability of techniques for forecasting infectious disease can make it easier to fight against infectious disease such as COVID-19.

## Forecasting COVID-19 spreading through an ensemble of classical and machine learning models: Spain’s case study [9]

This article combines both machine learning (ML) and classical population models, using exclusively publicly available data of incidence, mobility, vaccination and weather in Spain.

In this work the performance of four ML models were evaluated (Random Forest, Gradient Boosting, k-Nearest Neighbors and Kernel Ridge Regression), and four population models (Gompertz, Logistic, Richards and Bertalanffy) in order to estimate the near future evolution of the COVID-19 pandemic, using daily cases data, together with vaccination, mobility and weather data.

## Conclusion

The COVID-19 pandemic affected everyone around the world. It has made the researchers and investigation communities in different fields to find options to control the spread in record time. I would like to have the opportunity to duplicate some of the models to replicate some of the research that some communities have generated.

# Exploration Data Analysis

## Dataset

The dataset, provided by Our World in Data, provides COVID-19 vaccination information collected by **Our World in Data** available to **Kaggle community** <https://www.kaggle.com/datasets/caesarmario/our-world-in-data-covid19-dataset/download?datasetVersionNumber=418>. This dataset is updated daily, for the purpose of this study I am analyzing the data with information up to Oct 7th, 2023.

The dataset is a csv file, comma delimited with 67 variables and 346,567 observations.

## Data Dictionary

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 346567 entries, 0 to 346566

Data columns (total 67 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 iso\_code 346567 non-null object

1 continent 330089 non-null object

2 location 346567 non-null object

3 date 346567 non-null object

4 total\_cases 308672 non-null float64

5 new\_cases 337028 non-null float64

6 new\_cases\_smoothed 335769 non-null float64

7 total\_deaths 287169 non-null float64

8 new\_deaths 337072 non-null float64

9 new\_deaths\_smoothed 335842 non-null float64

10 total\_cases\_per\_million 308672 non-null float64

11 new\_cases\_per\_million 337028 non-null float64

12 new\_cases\_smoothed\_per\_million 335769 non-null float64

13 total\_deaths\_per\_million 287169 non-null float64

14 new\_deaths\_per\_million 337072 non-null float64

15 new\_deaths\_smoothed\_per\_million 335842 non-null float64

16 reproduction\_rate 184817 non-null float64

17 icu\_patients 37509 non-null float64

18 icu\_patients\_per\_million 37509 non-null float64

19 hosp\_patients 38759 non-null float64

20 hosp\_patients\_per\_million 38759 non-null float64

21 weekly\_icu\_admissions 10160 non-null float64

22 weekly\_icu\_admissions\_per\_million 10160 non-null float64

23 weekly\_hosp\_admissions 23145 non-null float64

24 weekly\_hosp\_admissions\_per\_million 23145 non-null float64

25 total\_tests 79387 non-null float64

26 new\_tests 75403 non-null float64

27 total\_tests\_per\_thousand 79387 non-null float64

28 new\_tests\_per\_thousand 75403 non-null float64

29 new\_tests\_smoothed 103965 non-null float64

30 new\_tests\_smoothed\_per\_thousand 103965 non-null float64

31 positive\_rate 95927 non-null float64

32 tests\_per\_case 94348 non-null float64

33 tests\_units 106788 non-null object

34 total\_vaccinations 78953 non-null float64

35 people\_vaccinated 75575 non-null float64

36 people\_fully\_vaccinated 72224 non-null float64

37 total\_boosters 47234 non-null float64

38 new\_vaccinations 65019 non-null float64

39 new\_vaccinations\_smoothed 180079 non-null float64

40 total\_vaccinations\_per\_hundred 78953 non-null float64

41 people\_vaccinated\_per\_hundred 75575 non-null float64

42 people\_fully\_vaccinated\_per\_hundred 72224 non-null float64

43 total\_boosters\_per\_hundred 47234 non-null float64

44 new\_vaccinations\_smoothed\_per\_million 180079 non-null float64

45 new\_people\_vaccinated\_smoothed 179887 non-null float64

46 new\_people\_vaccinated\_smoothed\_per\_hundred 179887 non-null float64

47 stringency\_index 197651 non-null float64

48 population\_density 294167 non-null float64

49 median\_age 273580 non-null float64

50 aged\_65\_older 264005 non-null float64

51 aged\_70\_older 270838 non-null float64

52 gdp\_per\_capita 268118 non-null float64

53 extreme\_poverty 172778 non-null float64

54 cardiovasc\_death\_rate 268731 non-null float64

55 diabetes\_prevalence 282404 non-null float64

56 female\_smokers 201575 non-null float64

57 male\_smokers 198833 non-null float64

58 handwashing\_facilities 131627 non-null float64

59 hospital\_beds\_per\_thousand 237221 non-null float64

60 life\_expectancy 318823 non-null float64

61 human\_development\_index 260466 non-null float64

62 population 346567 non-null float64

63 excess\_mortality\_cumulative\_absolute 11953 non-null float64

64 excess\_mortality\_cumulative 11953 non-null float64

65 excess\_mortality 11953 non-null float64

66 excess\_mortality\_cumulative\_per\_million 11953 non-null float64

dtypes: float64(62), object(5)

memory usage: 177.2+ MB

## Metadata

The dataset size is 91.1 MB, the Pandas data profiling is almost 300 MB. This initial analysis can be found in GitHub <https://github.com/aamadorc/CIND820/commits/main/CIND820_EDA.html> under version “Pandas Data Profiling” - 759d19f. Due to the size of the file, on GitHub is loaded with Git LFS which means it can not be displayed but available to be downloaded.

Below is the metadata of the whole dataset.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| iso\_code | ISO 3166-1 alpha-3 – three-letter country codes. Note that OWID-defined regions (e.g. continents like 'Europe') contain prefix 'OWID\_'. |
| continent | Continent of the geographical location |
| location | Geographical location |
| date | Date of observation |
| total\_cases | Total confirmed cases of COVID-19. Counts can include probable cases, where reported. |
| new\_cases | New confirmed cases of COVID-19. Counts can include probable cases, where reported. In rare cases where our source reports a negative daily change due to a data correction, we set this metric to NA. |
| new\_cases\_smoothed | New confirmed cases of COVID-19 (7-day smoothed). Counts can include probable cases, where reported. |
| total\_deaths | Total deaths attributed to COVID-19. Counts can include probable deaths, where reported. |
| new\_deaths | New deaths attributed to COVID-19. Counts can include probable deaths, where reported. In rare cases where our source reports a negative daily change due to a data correction, we set this metric to NA. |
| new\_deaths\_smoothed | New deaths attributed to COVID-19 (7-day smoothed). Counts can include probable deaths, where reported. |
| total\_cases\_per\_million | Total confirmed cases of COVID-19 per 1,000,000 people. Counts can include probable cases, where reported. |
| new\_cases\_per\_million | New confirmed cases of COVID-19 per 1,000,000 people. Counts can include probable cases, where reported. |
| new\_cases\_smoothed\_per\_million | New confirmed cases of COVID-19 (7-day smoothed) per 1,000,000 people. Counts can include probable cases, where reported. |
| total\_deaths\_per\_million | Total deaths attributed to COVID-19 per 1,000,000 people. Counts can include probable deaths, where reported. |
| new\_deaths\_per\_million | New deaths attributed to COVID-19 per 1,000,000 people. Counts can include probable deaths, where reported. |
| new\_deaths\_smoothed\_per\_million | New deaths attributed to COVID-19 (7-day smoothed) per 1,000,000 people. Counts can include probable deaths, where reported. |
| reproduction\_rate | Real-time estimate of the effective reproduction rate (R) of COVID-19. See https://github.com/crondonm/TrackingR/tree/main/Estimates-Database |
| icu\_patients | Number of COVID-19 patients in intensive care units (ICUs) on a given day |
| icu\_patients\_per\_million | Number of COVID-19 patients in intensive care units (ICUs) on a given day per 1,000,000 people |
| hosp\_patients | Number of COVID-19 patients in hospital on a given day |
| hosp\_patients\_per\_million | Number of COVID-19 patients in hospital on a given day per 1,000,000 people |
| weekly\_icu\_admissions | Number of COVID-19 patients newly admitted to intensive care units (ICUs) in a given week (reporting date and the preceeding 6 days) |
| weekly\_icu\_admissions\_per\_million | Number of COVID-19 patients newly admitted to intensive care units (ICUs) in a given week per 1,000,000 people (reporting date and the preceeding 6 days) |
| weekly\_hosp\_admissions | Number of COVID-19 patients newly admitted to hospitals in a given week (reporting date and the preceeding 6 days) |
| weekly\_hosp\_admissions\_per\_million | Number of COVID-19 patients newly admitted to hospitals in a given week per 1,000,000 people (reporting date and the preceeding 6 days) |
| total\_tests | Total tests for COVID-19 |
| new\_tests | New tests for COVID-19 (only calculated for consecutive days) |
| total\_tests\_per\_thousand | Total tests for COVID-19 per 1,000 people |
| new\_tests\_per\_thousand | New tests for COVID-19 per 1,000 people |
| new\_tests\_smoothed | New tests for COVID-19 (7-day smoothed). For countries that don't report testing data on a daily basis, we assume that testing changed equally on a daily basis over any periods in which no data was reported. This produces a complete series of daily figures, which is then averaged over a rolling 7-day window |
| new\_tests\_smoothed\_per\_thousand | New tests for COVID-19 (7-day smoothed) per 1,000 people |
| positive\_rate | The share of COVID-19 tests that are positive, given as a rolling 7-day average (this is the inverse of tests\_per\_case) |
| tests\_per\_case | Tests conducted per new confirmed case of COVID-19, given as a rolling 7-day average (this is the inverse of positive\_rate) |
| tests\_units | Units used by the location to report its testing data. A country file can't contain mixed units. All metrics concerning testing data use the specified test unit. Valid units are 'people tested' (number of people tested), 'tests performed' (number of tests performed. a single person can be tested more than once in a given day) and 'samples tested' (number of samples tested. In some cases, more than one sample may be required to perform a given test.) |
| total\_vaccinations | Total number of COVID-19 vaccination doses administered |
| people\_vaccinated | Total number of people who received at least one vaccine dose |
| people\_fully\_vaccinated | Total number of people who received all doses prescribed by the initial vaccination protocol |
| total\_boosters | Total number of COVID-19 vaccination booster doses administered (doses administered beyond the number prescribed by the vaccination protocol) |
| new\_vaccinations | New COVID-19 vaccination doses administered (only calculated for consecutive days) |
| new\_vaccinations\_smoothed | New COVID-19 vaccination doses administered (7-day smoothed). For countries that don't report vaccination data on a daily basis, we assume that vaccination changed equally on a daily basis over any periods in which no data was reported. This produces a complete series of daily figures, which is then averaged over a rolling 7-day window |
| total\_vaccinations\_per\_hundred | Total number of COVID-19 vaccination doses administered per 100 people in the total population |
| people\_vaccinated\_per\_hundred | Total number of people who received at least one vaccine dose per 100 people in the total population |
| people\_fully\_vaccinated\_per\_hundred | Total number of people who received all doses prescribed by the initial vaccination protocol per 100 people in the total population |
| total\_boosters\_per\_hundred | Total number of COVID-19 vaccination booster doses administered per 100 people in the total population |
| new\_vaccinations\_smoothed\_per\_million | New COVID-19 vaccination doses administered (7-day smoothed) per 1,000,000 people in the total population |
| new\_people\_vaccinated\_smoothed | Daily number of people receiving their first vaccine dose (7-day smoothed) |
| new\_people\_vaccinated\_smoothed\_per\_hundred | Daily number of people receiving their first vaccine dose (7-day smoothed) per 100 people in the total population |
| stringency\_index | Government Response Stringency Index: composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response) |
| population\_density | Number of people divided by land area, measured in square kilometers, most recent year available |
| median\_age | Median age of the population, UN projection for 2020 |
| aged\_65\_older | Share of the population that is 65 years and older, most recent year available |
| aged\_70\_older | Share of the population that is 70 years and older in 2015 |
| gdp\_per\_capita | Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available |
| extreme\_poverty | Share of the population living in extreme poverty, most recent year available since 2010 |
| cardiovasc\_death\_rate | Death rate from cardiovascular disease in 2017 (annual number of deaths per 100,000 people) |
| diabetes\_prevalence | Diabetes prevalence (% of population aged 20 to 79) in 2017 |
| female\_smokers | Share of women who smoke, most recent year available |
| male\_smokers | Share of men who smoke, most recent year available |
| handwashing\_facilities | Share of the population with basic handwashing facilities on premises, most recent year available |
| hospital\_beds\_per\_thousand | Hospital beds per 1,000 people, most recent year available since 2010 |
| life\_expectancy | Life expectancy at birth in 2019 |
| human\_development\_index | A composite index measuring average achievement in three basic dimensions of human development—a long and healthy life, knowledge and a decent standard of living. Values for 2019, imported from http://hdr.undp.org/en/indicators/137506 |
| population | Population (latest available values). See https://github.com/owid/covid-19-data/blob/master/scripts/input/un/population\_latest.csv for full list of sources |
| excess\_mortality\_cumulative\_absolute | Cumulative difference between the reported number of deaths since 1 January 2020 and the projected number of deaths for the same period based on previous years. For more information, see https://github.com/owid/covid-19-data/tree/master/public/data/excess\_mortality |
| excess\_mortality\_cumulative | Percentage difference between the cumulative number of deaths since 1 January 2020 and the cumulative projected deaths for the same period based on previous years. For more information, see https://github.com/owid/covid-19-data/tree/master/public/data/excess\_mortality |
| excess\_mortality | Percentage difference between the reported number of weekly or monthly deaths in 2020–2021 and the projected number of deaths for the same period based on previous years. For more information, see https://github.com/owid/covid-19-data/tree/master/public/data/excess\_mortality |
| excess\_mortality\_cumulative\_per\_million | Cumulative difference between the reported number of deaths since 1 January 2020 and the projected number of deaths for the same period based on previous years, per million people. For more information, see https://github.com/owid/covid-19-data/tree/master/public/data/excess\_mortality |

Due to the size of the file, I had implemented different techniques to manage the number of dimensions and size.

## Dimensionality Reduction (CMTH642 - Module 9)

There are seven techniques for Dimensionality Reduction: Missing Values, Low Variance Filter, High Correlation Filter, PCA, Random Forests, Backward Feature Elimination, and Forward Feature Construction. [14]

Before starting with dimensionality reduction techniques, we can see on the dataset information since the beginning of the pandemic. Because the goal of this study is the prediction of total cases and this was affected by the time the vaccine became available, data before the vaccine will be removed. Multiple vaccines became available during the second semester of 2020. By December most countries have approved vaccines, then data before Jan 1st, 2021 was removed.

In addition, after verifying the categorical attributes iso\_code and location, the conclusion is that one can be derived from the other. For the purpose of this study, location was kept and iso\_code removed.

### Removing data columns with too many NaN values

We can calculate the ratio of missing values using a simple formula. The formula is: the number of missing values in each column divided by the total number of observations. Generally, we can drop variables having a missing value ratio of more than 60% or 70%. For the purpose of the study the threshold is 60% of missing values and those attributes will be removed.

### Low Variance Filter

Another way of measuring how much information a data column has, is to measure its variance. In the limit case where the column cells assume a constant value, the variance would be 0 and the column would be of no help in the discrimination of different groups of data.

The Low Variance Filter calculates each column variance and removes those columns with a variance value below a given threshold. Variance can only be calculated for numerical columns, i.e. this dimensionality reduction method applies only to numerical columns. Note, too, that the variance value depends on the column numerical range. Therefore data column ranges need to be normalized to make variance values independent from the column domain range.

First a Normalizer normalizes all column ranges to [0, 1]; next, a Low Variance Filter calculates the columns variance and filters out the columns with a variance lower than a set threshold.

After applying this technique, no dimensions were lower than the set threshold, but 3 dimensions had a variance of 0 and were removed.

### High correlation with other data columns

Features with high correlation are more linearly dependent and hence have almost the same effect on the dependent variable. So, when two features have high correlation, one of the two features can be dropped.

In addition, removal of different features from the dataset will have different effects on the p-value for the dataset. We can remove different features and measure the p-value in each case. These measured p-values can be used to decide whether to keep a feature or not. p-value used fot the study is 0.05.

### Summary of the techniques

|  |  |  |  |
| --- | --- | --- | --- |
| **Technique** | **# Observations** | **# Attributes** | **Size** |
| Raw | 346,566 | 67 | 23,219,989 |
| Data before 2022-01-01 removed | 255,173 | 67 | 17,096,591 |
| iso\_code removed | 255,173 | 66 | 16,841,418 |
| 60% of missing values removed | 255,173 | 35 | 8,931,055 |
| Low Variance Filter | 255,173 | 32 | 8,165,536 |
| High correlation with other data columns | 255,173 | 26 | 6,634,498 |

## Dataset subset

Metadata of the subset, 26 variables, and 255,173 observations:

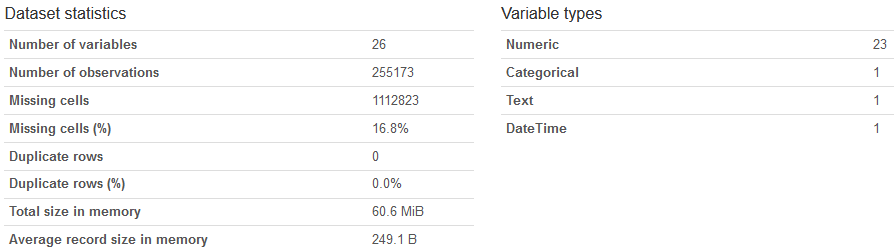
|  |  |
| --- | --- |
| **Variable** | **Description** |
| continent | Continent of the geographical location |
| location | Geographical location |
| date | Date of observation |
| total\_cases | Total confirmed cases of COVID-19. Counts can include probable cases, where reported. |
| new\_cases | New confirmed cases of COVID-19. Counts can include probable cases, where reported. In rare cases where our source reports a negative daily change due to a data correction, we set this metric to NA. |
| new\_cases\_smoothed | New confirmed cases of COVID-19 (7-day smoothed). Counts can include probable cases, where reported. |
| new\_deaths\_smoothed | New deaths attributed to COVID-19 (7-day smoothed). Counts can include probable deaths, where reported. |
| total\_cases\_per\_million | Total confirmed cases of COVID-19 per 1,000,000 people. Counts can include probable cases, where reported. |
| new\_cases\_smoothed\_per\_million | New confirmed cases of COVID-19 (7-day smoothed) per 1,000,000 people. Counts can include probable cases, where reported. |
| total\_deaths\_per\_million | Total deaths attributed to COVID-19 per 1,000,000 people. Counts can include probable deaths, where reported. |
| new\_deaths\_smoothed\_per\_million | New deaths attributed to COVID-19 (7-day smoothed) per 1,000,000 people. Counts can include probable deaths, where reported. |
| new\_vaccinations\_smoothed\_per\_million | New COVID-19 vaccination doses administered (7-day smoothed) per 1,000,000 people in the total population |
| new\_people\_vaccinated\_smoothed | Daily number of people receiving their first vaccine dose (7-day smoothed) |
| stringency\_index | Government Response Stringency Index: composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response) |
| population\_density | Number of people divided by land area, measured in square kilometers, most recent year available |
| aged\_65\_older | Share of the population that is 65 years and older, most recent year available |
| gdp\_per\_capita | Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available |
| extreme\_poverty | Share of the population living in extreme poverty, most recent year available since 2010 |
| cardiovasc\_death\_rate | Death rate from cardiovascular disease in 2017 (annual number of deaths per 100,000 people) |
| diabetes\_prevalence | Diabetes prevalence (% of population aged 20 to 79) in 2017 |
| female\_smokers | Share of women who smoke, most recent year available |
| male\_smokers | Share of men who smoke, most recent year available |
| life\_expectancy | Life expectancy at birth in 2019 |
| population | Population (latest available values). See https://github.com/owid/covid-19-data/blob/master/scripts/input/un/population\_latest.csv for full list of sources |
| hospital\_beds\_per\_thousand | Hospital beds per 1,000 people, most recent year available since 2010 |
| total\_deaths | Total deaths attributed to COVID-19. Counts can include probable deaths, where reported. |

After analyzing the attributes that were eliminated and kept, data like testing doesn’t add value in the prediction of the number of cases. As well as derived columns like weekly columns are not relevant. ICU information doesn't affect the number of cases either.

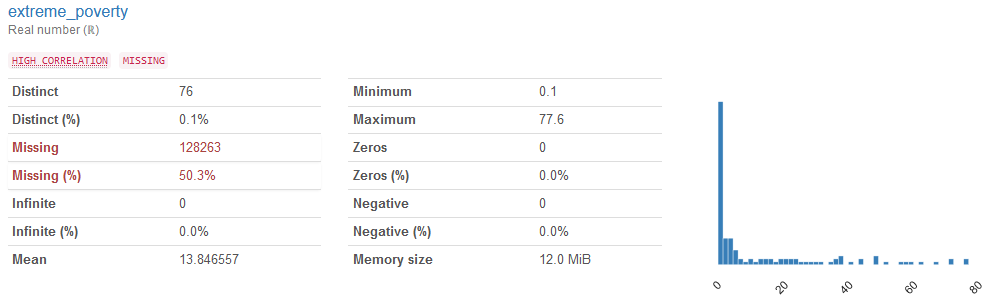
Also, it is interesting to see how information like gross domestic product at purchasing power, and extreme poverty are pretty relevant to influence the number of cases. Based on the information the age range that is most affected is people who are 65 and older. And surprisingly, the number of hospital beds does influence the number of cases.

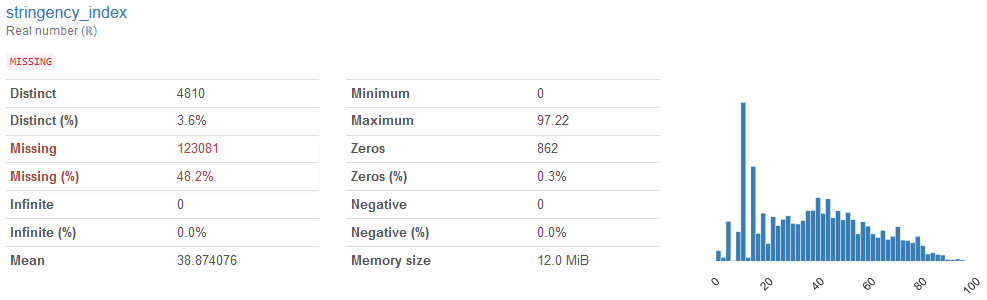
The categorical attributes are: date, location, and continent.

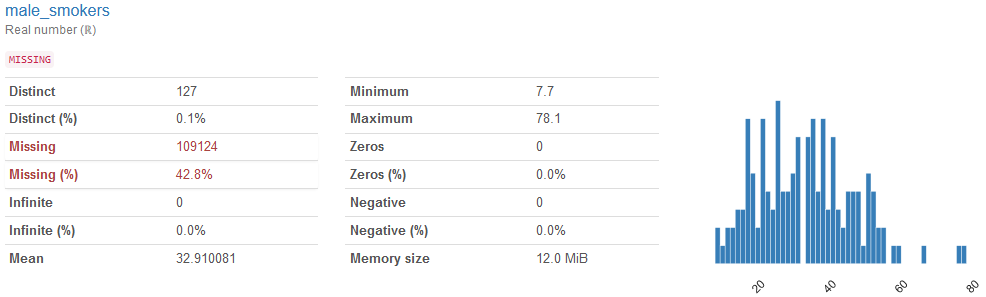
#### Overview

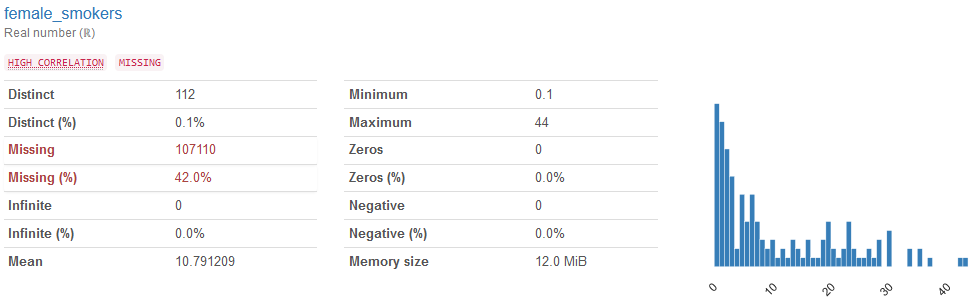


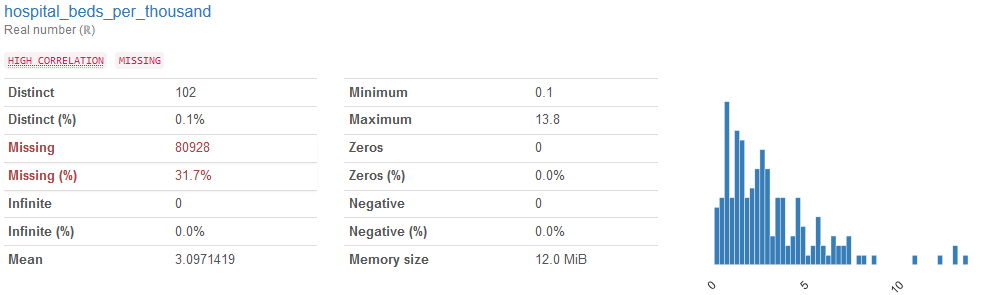
#### Variables

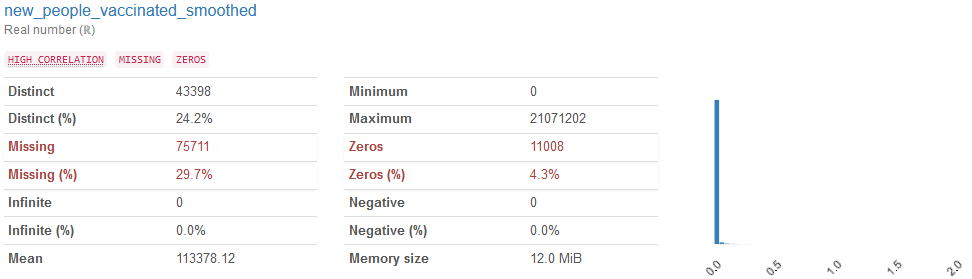


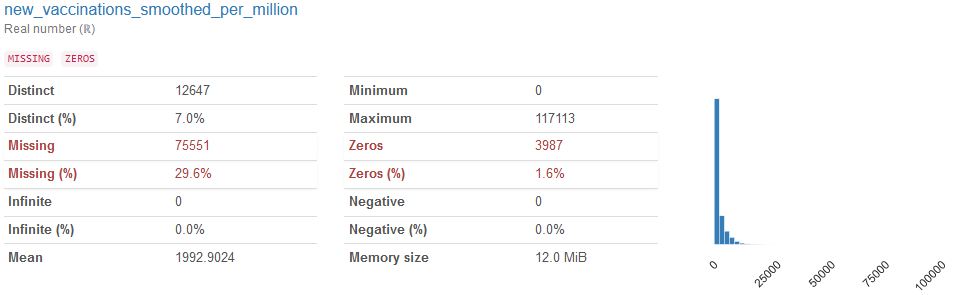


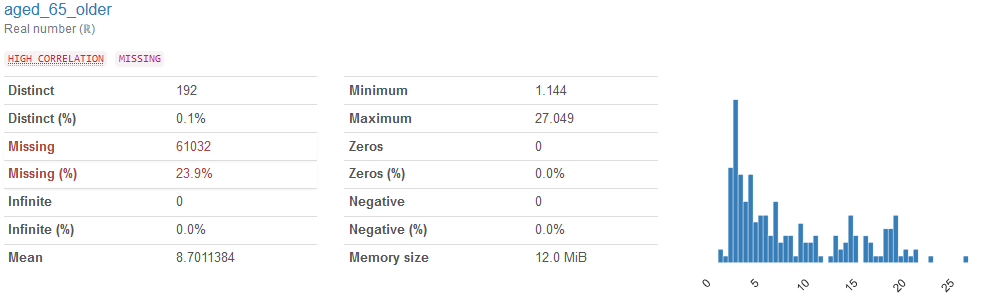


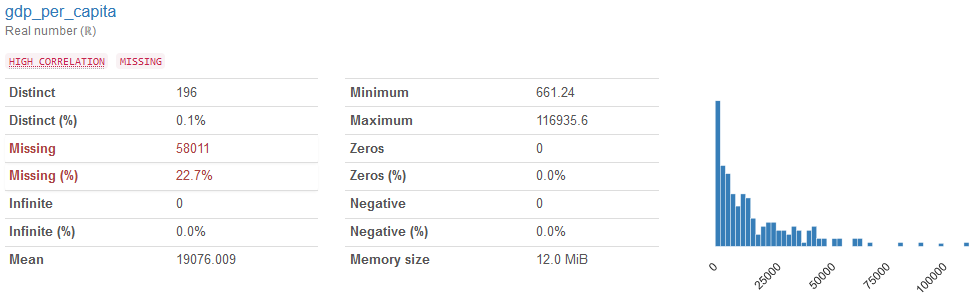


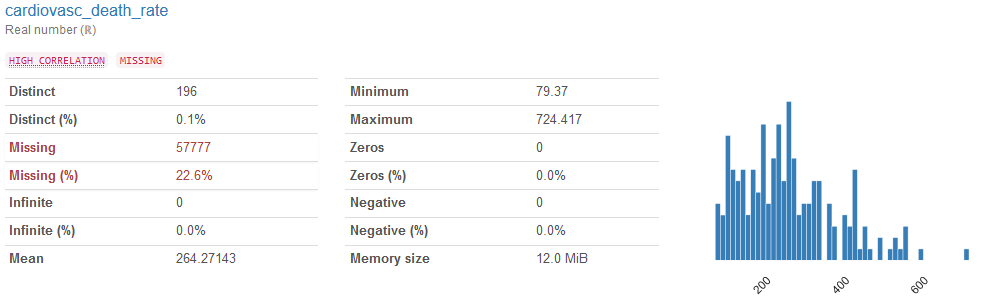


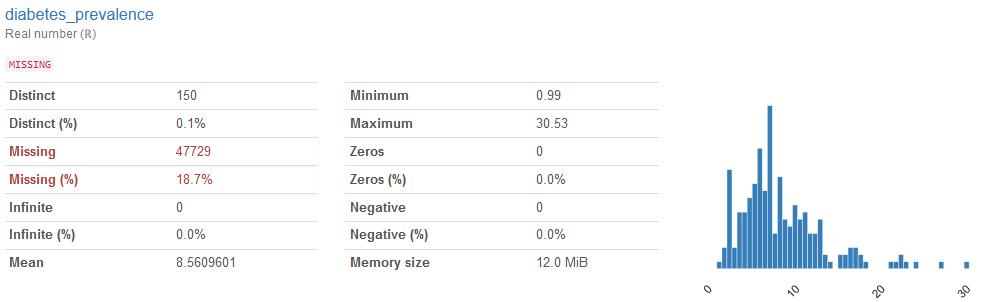


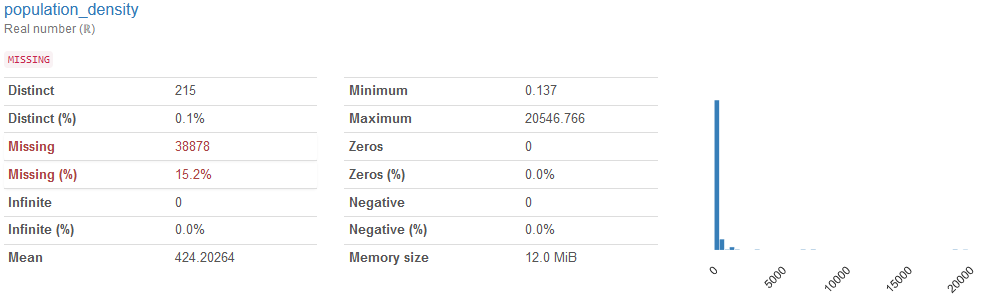


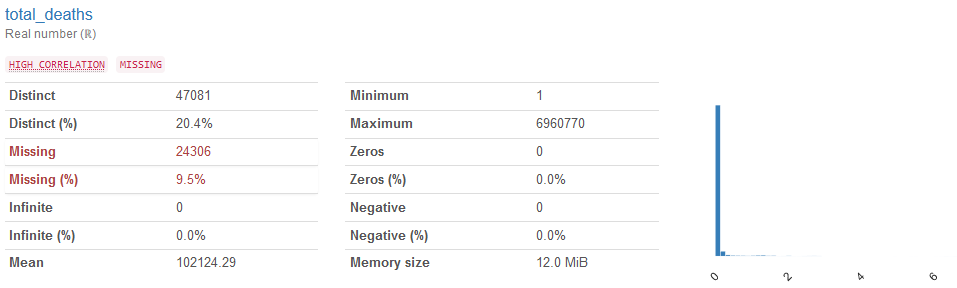


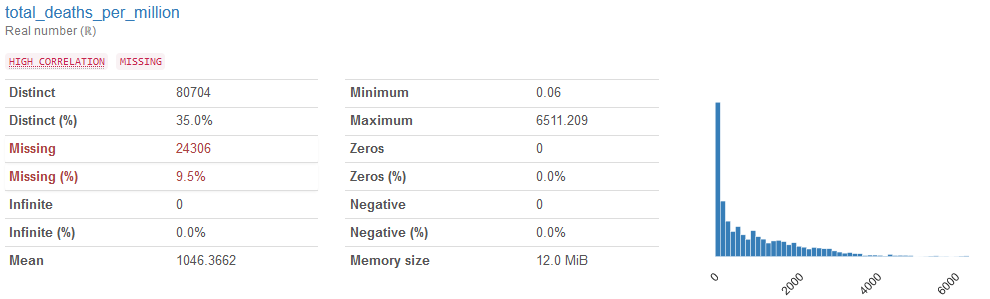


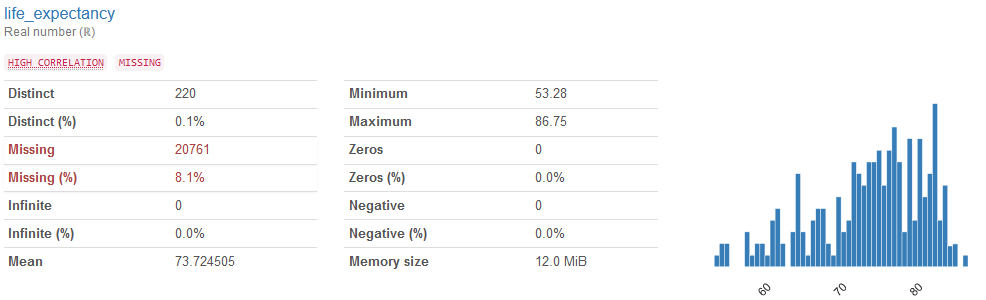


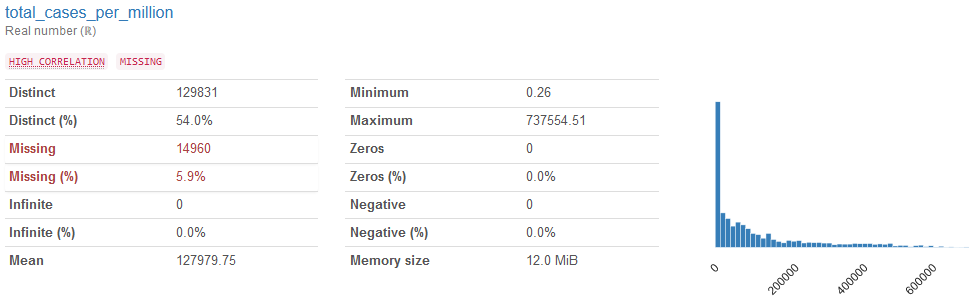


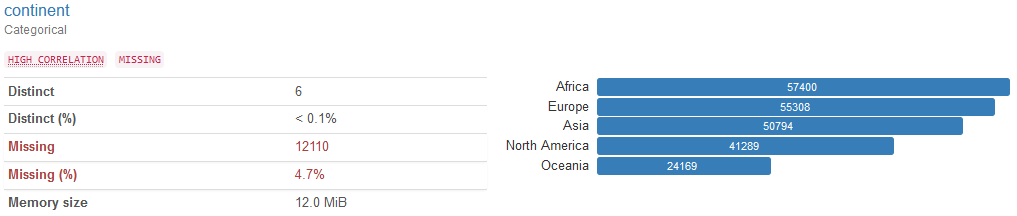


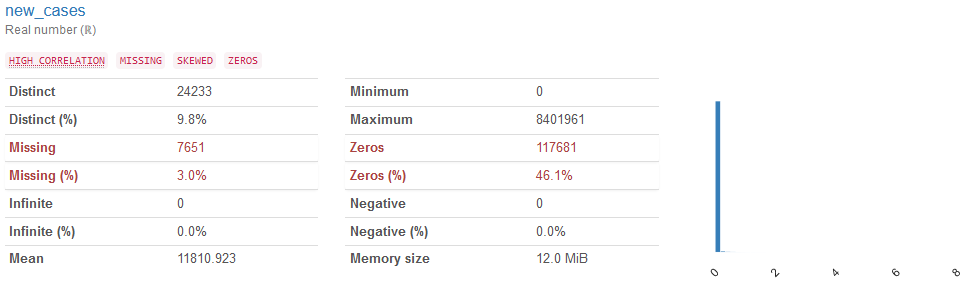


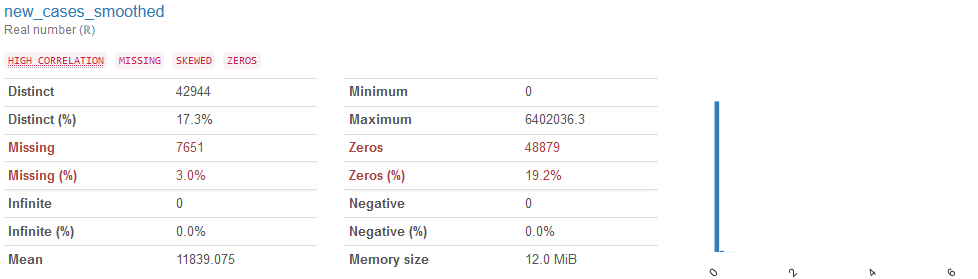


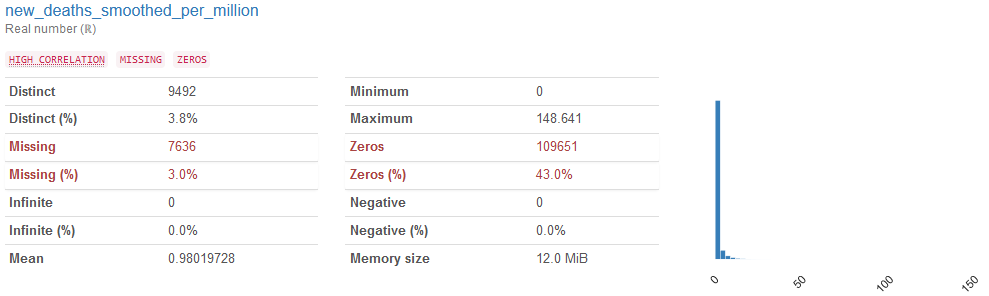


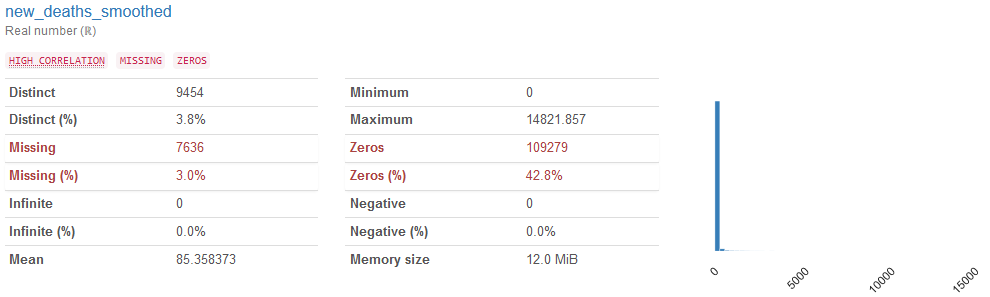


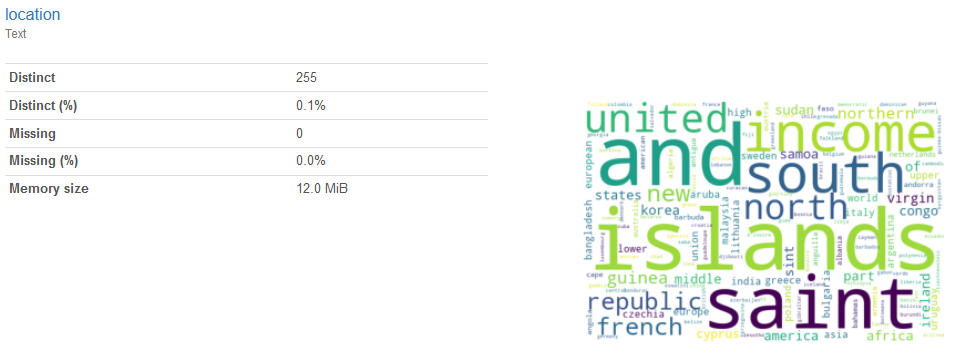


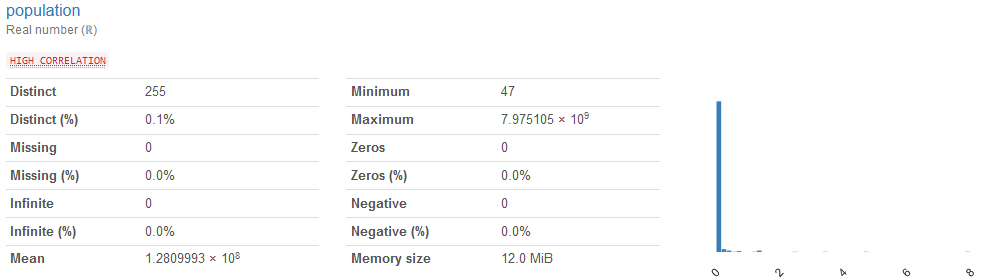


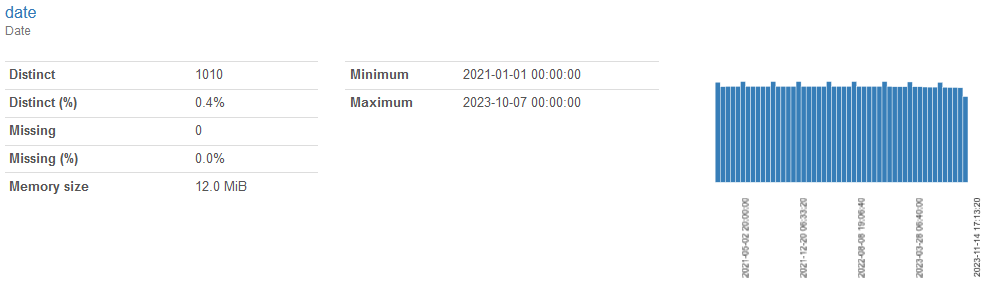


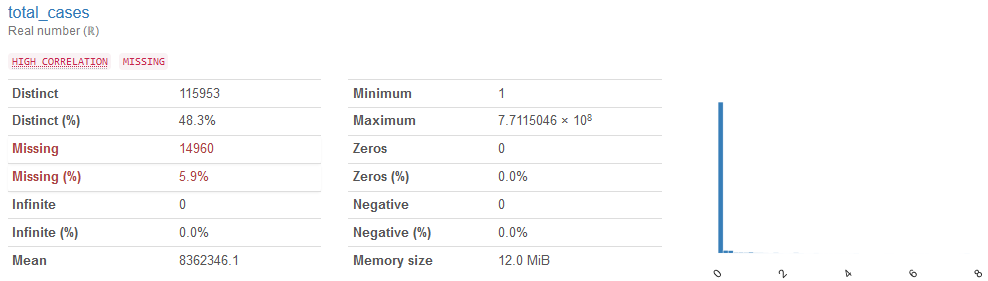


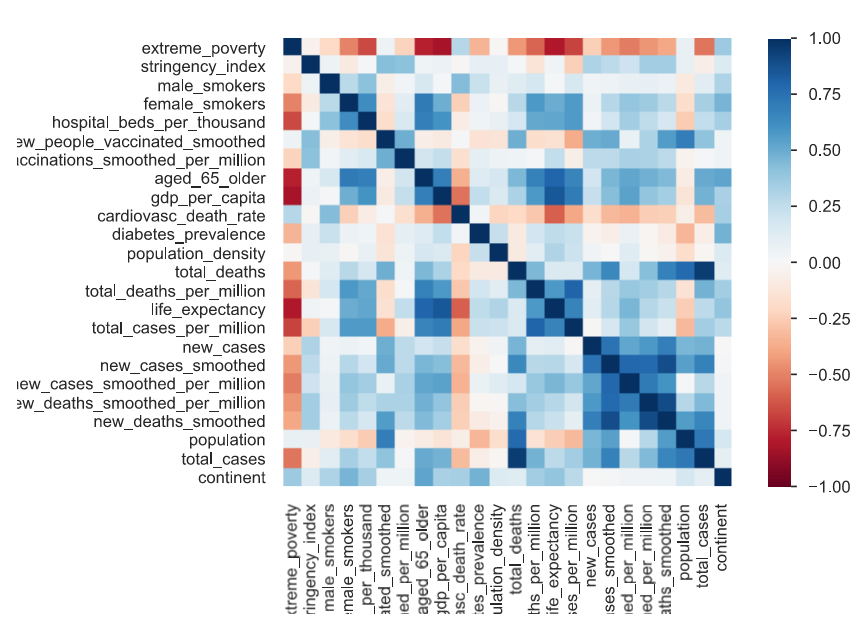












## GITHUB R**epository**:

https://github.com/aamadorc/CIND820

## References

1. Ahouz, F., Golabpour, A. Predicting the incidence of COVID-19 using data mining. *BMC Public Health* **21**, 1087 (2021). https://doi.org/10.1186/s12889-021-11058-3
2. Gothai E, Thamilselvan R, Rajalaxmi RR, Sadana RM, Ragavi A, Sakthivel R. Prediction of COVID-19 growth and trend using machine learning approach. Mater Today Proc. 2023;81:597-601. doi: 10.1016/j.matpr.2021.04.051. Epub 2021 Apr 15. PMID: 33880331; PMCID: PMC8049379.
3. Zoabi, Y., Deri-Rozov, S. & Shomron, N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *npj Digit. Med.* **4**, 3 (2021). https://doi.org/10.1038/s41746-020-00372-6
4. Meraihi Y, Gabis AB, Mirjalili S, Ramdane-Cherif A, Alsaadi FE. Machine Learning-Based Research for COVID-19 Detection, Diagnosis, and Prediction: A Survey. SN Comput Sci. 2022;3(4):286. doi: 10.1007/s42979-022-01184-z. Epub 2022 May 12. PMID: 35578678; PMCID: PMC9096341.
5. ????L. Wang, H. Shen, K. Enfield and K. Rheuban, "COVID-19 Infection Detection Using Machine Learning," *2021 IEEE International Conference on Big Data (Big Data)*, Orlando, FL, USA, 2021, pp. 4780-4789, doi: 10.1109/BigData52589.2021.9671700.
6. Painuli D, Mishra D, Bhardwaj S, Aggarwal M. Forecast and prediction of COVID-19 using machine learning. Data Science for COVID-19. 2021:381–97. doi: 10.1016/B978-0-12-824536-1.00027-7. Epub 2021 May 21. PMCID: PMC8138040.
7. Gangloff, C., Rafi, S., Bouzillé, G. *et al.* Machine learning is the key to diagnose COVID-19: a proof-of-concept study. *Sci Rep* **11**, 7166 (2021). https://doi.org/10.1038/s41598-021-86735-9
8. Tuli, Shreshth & Tuli, Shikhar & Tuli, Rakesh & Gill, Sukhpal Singh. (2020). Predicting the Growth and Trend of COVID-19 Pandemic using Machine Learning and Cloud Computing. 100222. 10.1016/j.iot.2020.100222.
9. Heredia Cacha, I., Sáinz-Pardo Díaz, J., Castrillo, M. *et al.* Forecasting COVID-19 spreading through an ensemble of classical and machine learning models: Spain’s case study. *Sci Rep* **13**, 6750 (2023). https://doi.org/10.1038/s41598-023-33795-8
10. Barua, Arpita and Hridoy, Monowar Wadud and Uddin, Kazi Riad and Chowdhury, Ratri and Ahamed, Jamal Uddin, Analysis and Prediction of the Spread of COVID-19 in Bangladesh Using Statistical and Machine Learning Approach. Available at SSRN: <https://ssrn.com/abstract=4592228> or [http://dx.doi.org/10.2139/ssrn.4592228](https://dx.doi.org/10.2139/ssrn.4592228)
11. Saqib, Mohd. (2021). Forecasting COVID-19 Outbreak Progression Using Hybrid Polynomial-Bayesian Ridge Regression Model. Applied Intelligence. 10.1007/s10489-020-01942-7.
12. Iwendi, Celestine & Bashir, Ali & Pasupuleti, Naga & Radha, Suja & Chatterjee, Jyotir & Peshkar, Atharva & Mishra, Rishita & Pillai, Sofia & Jo, Ohyun. (2020). COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. Frontiers in Public Health. 8. 10.3389/fpubh.2020.00357.
13. Singh KK, Kumar S, Dixit P, Bajpai MK. Kalman filter based short term prediction model for COVID-19 spread. Appl Intell (Dordr). 2021;51(5):2714-2726. doi: 10.1007/s10489-020-01948-1. Epub 2020 Nov 3. PMID: 34764569; PMCID: PMC7676285.
14. Silipo, R., Adae, I., Hart, A., & Berthold, M. (2014). Seven techniques for dimensionality reduction [PDF]. KNIME. Retrieved from <https://files.knime.com/sites/default/files/inline-images/knime_seventechniquesdatadimreduction.pdf>
15. <https://www.analyticsvidhya.com/blog/2021/04/beginners-guide-to-missing-value-ratio-and-its-implementation/>
16. <https://www.kaggle.com/code/bbloggsbott/feature-selection-correlation-and-p-value>