

COVID-19: BEHIND THE NUMBERS

Data Mining and Machine Learning Project

Prof. Marcelloni Francesco

Prof. Ducange Pietro

Rambod Rahmani

Master's Degree in Artificial Intelligence and Data Engineering

February 8, 2021

Contents

1	Introduction	1
2	Dataset	2
3	Analysis	2
3.1	Which countries have been affected the most by COVID-19?	2
3.2	Which governments have taken the right actions?	3
3.3	Personalized predictive models for symptomatic COVID-19 patients	3
4	Conclusion	3
5	Software Architecture	3

1 Introduction

Decemebr 31, 2019: *Wuhan Municipal Health Commission, China, reported a cluster of cases of pneumonia in Wuhan, Hubei Province. A novel coronavirus was eventually identified.*

January 1, 2020: *World Health Organization (WHO) had set up the Incident Management Support Team across the three levels of the organization: headquarters, regional headquarters and country level, putting the organization on an emergency footing for dealing with the outbreak.*

January 5, 2020: *WHO published the first Disease Outbreak News on the new virus. This was a flagship technical publication to the scientific and public health community as well as global media.*

January 12, 2020: *China publicly shared the genetic sequence of COVID-19.*

At the beginning of 2020, a new virus started spreading around in the capital of Central China's Hubei province: the city we all came to know as Wuhan. As it turned out, this was the start of a world-changing event with overwhelming extent: Coronavirus Disease 2019 (COVID-19). After the first wave of the virus has passed over the entire world, the aim of this work is to address the following questions:

- Which countries have been affected the most by COVID-19?
- Which governments have taken the right actions to stop the spreading of the virus?
- Is it possible to build personalized predictive models for symptomatic COVID-19 patients based on basic health preconditions?

In order to fully answer these questions, first of all a reliable and big enough dataset was needed. Second, Data Mining and Machine Learning techniques were applied in order to obtain statistically significant results that could help address the proposed questions. In the following pages the described work and the resulting Python software is presented. The software architecture is presented in the very last section in order to focus primarily on the dataset retrieval and preprocessing, and on the analysis techniques and results.

2 Dataset

3 Analysis

As said in the introductory section, the analysis was carried out using data mining and machine learning techniques in order to answer the proposed questions. Each of the following subsections is focused on one of them.

3.1 Which countries have been affected the most by COVID-19?

To answer the very first question, we need to understand what is hidden behind the official numbers and charts of confirmed COVID-19 active cases and deaths. We are so used to watching them and sometimes we think we might even understand how the COVID-19 pandemic is evolving as days goes by. For example, it is very common to consider the following:

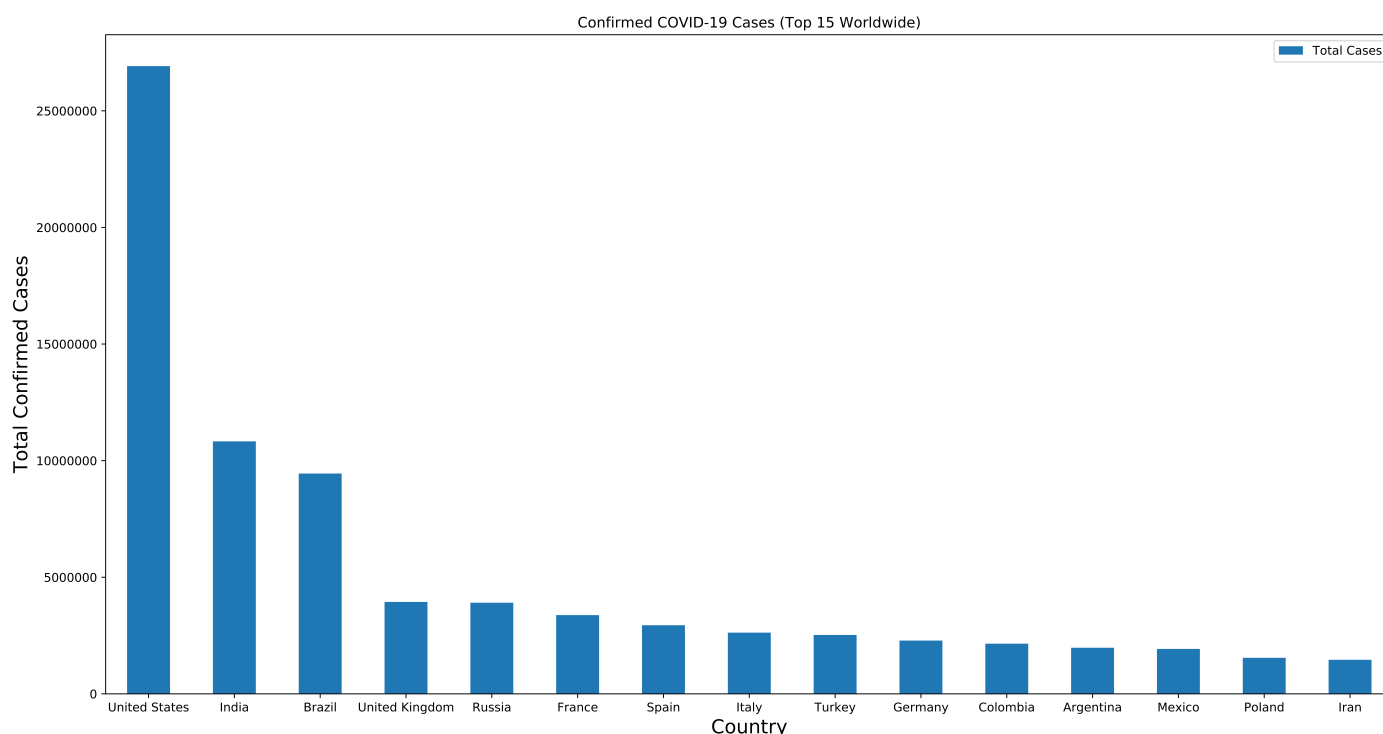


Figure 1: Confirmed COVID-19 Cases Worldwide

Is this really the right choice? Does this ranking tells us anything meaningful about the current undergoing pandemic situation? From what we can observe in Figure 1, clearly United States has higher confirmed COVID-19 cases than countries such as Spain or Italy. Taking into account that the US is a much bigger country, we can agree that this results do not imply that the US is more affected than Spain, Italy or Germany. We can therefore think of a fairer comparison independent of the country size: the number of infections needs to be normalized to the population of each country:

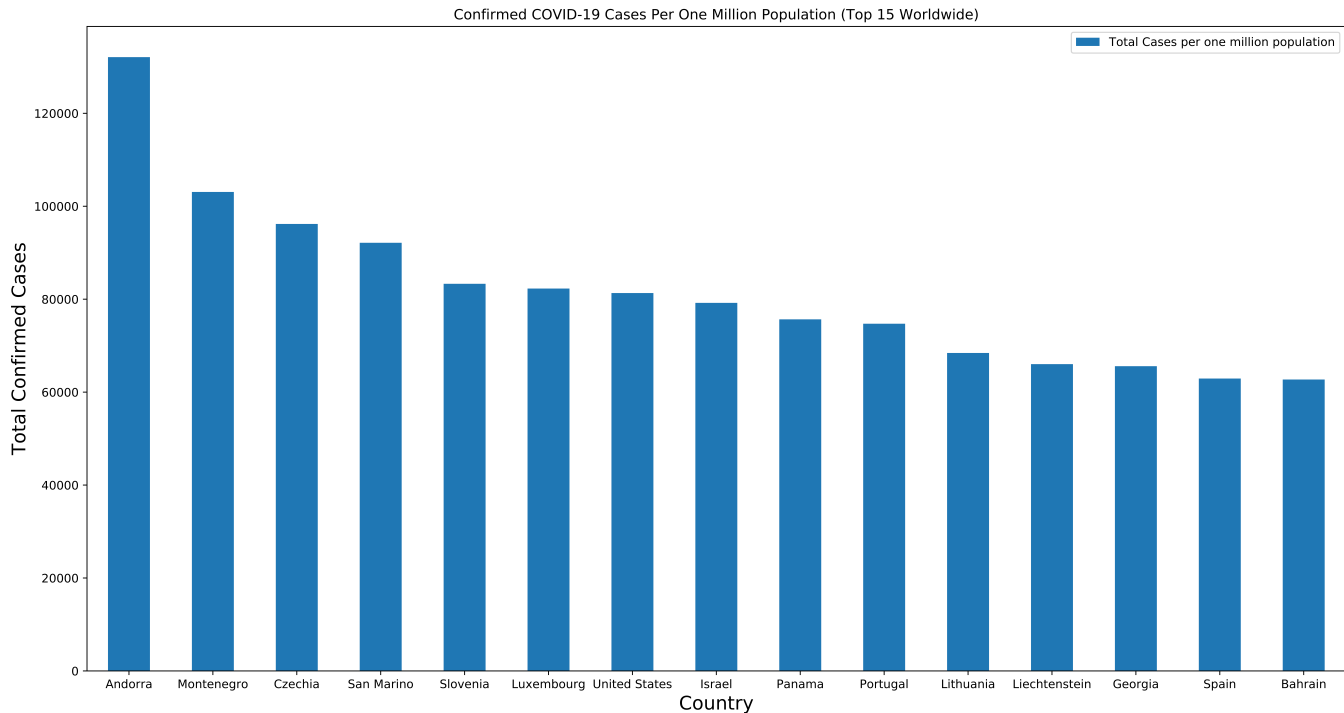


Figure 2: Confirmed COVID-19 Cases Per One Million Population Worldwide

This provides a more coherent view of the countries most affected by COVID-19, however it is not yet good enough: countries have different testing policies and a more intensive COVID-19 testing rate gives more confirmed cases while no testing at all would imply zero cases. We can all agree upon the fact that zero cases with no testing at all does not really mean that a given country is not affected by the pandemic. We therefore need a quantity unrelated to the rate of testing.

This quantity is the number of deaths: this value is unbiased by the testing rate. We will use the normalized number of deaths for comparing countries. To summarize, by means of a machine learning algorithm applied to the COVID-19 data we organized countries into groups with similar epidemiological behavior. Surprisingly, those groups form local clusters on the world map as well. This unexpected insight helps us to answer the question proposed in this section.

3.2 Which governments have taken the right actions?

3.3 Personalized predictive models for symptomatic COVID-19 patients

4 Conclusion

5 Software Architecture

References

- [1] World Health Organization. *Archived: WHO Timeline - COVID-19*. 2021. URL: <https://www.who.int/news/item/27-04-2020-who-timeline---covid-19>.
- [2] Towards Data Science - Robert Biele. *COVID-19: What Is Hidden Behind the Official Numbers?* 2020. URL: <https://towardsdatascience.com/which-countries-are-affected-the-most-by-covid-19-4d4570852e31>.
- [3] Our World in Data. *Coronavirus Pandemic (COVID-19) – the data*. 2021. URL: <https://ourworldindata.org/coronavirus-data>.
- [4] Ioannis Ch. Paschalidis Salomon Wollenstein-Betech Christos G. Cassandras. “Personalized Predictive Models for Symptomatic COVID-19 Patients Using Basic Preconditions”. In: *International Journal of Medical Informatics* 142 (2020). DOI: <https://doi.org/10.1016/j.ijmedinf.2020.104258>.
- [5] Gobierno de Mexico - Direccion General de Epidemiologia. *COVID-19 Datos Abiertos*. 2021. URL: <https://www.gob.mx/salud/documentos/datos-abiertos-152127>.