Data analysis of CoV-19 pandemic with visual interactive dashboard using CRISP-DM

**Contents**

1. [Introduction](#_bookmark0) 2
2. [Business Understanding](#_bookmark1) 3
   1. [Requirements](#_bookmark2) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3
   2. [Constraints](#_bookmark3) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3
   3. [Assumptions](#_bookmark4) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3
3. [Data Understanding](#_bookmark5) 4
   1. [Data Collection](#_bookmark6) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4
   2. [Description](#_bookmark8) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4
   3. [Exploratory Data Analysis (EDA)](#_bookmark10) . . . . . . . . . . . . . . . . . . . . . . . . 5
   4. [Data Quality Issues](#_bookmark15) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8
4. [Data Preparation](#_bookmark16) 8
   1. [Data Selection](#_bookmark17) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8
   2. [Data Transformation](#_bookmark19) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9
5. [Modelling](#_bookmark21) 10
6. [Evaluation](#_bookmark25) 13
7. [Deployment](#_bookmark27) 14
8. [Visualisations](#_bookmark29) 15
   1. [Existing Data Visualisation](#_bookmark30) 15
   2. [Predicted Data Visualisation](#_bookmark32) 17
9. [Conclusion](#_bookmark35) 18

### References

### Appendix (Source Code)

# Introduction

Society’s vulnerability to contain the spread of a highly infectious contagion was demon- started during 2002 in China, when the novel severe acute respiratory syndrome Coronavirus (SARS-CoV) emerged and spread throughout Asia and Canada. During this time, the spread of the SARS-CoV had largely been controlled and eliminated by means of quarantine procedures and infection preventative measures without a direct antiviral solution [.](#_bookmark40) In the city of Wuhan in mainland China, during the end of 2019, the pattern seems to have repeated itself with the emergence of a new type of the virus genetically related to SARS-CoV, referred to as CoV-19.

The sudden increase in confirmed CoV-19 cases in Wuhan and the subsequent spreading of the disease to other countries has sparked a sense of global alarm. Therefore, it is imperative that society at large be informed about the factual data to reaffirm the urgency of the pandemic while at the same time avoiding irrational panic.

This report intends to provide a method to inform individuals of the circumstances of the pandemic by means of an interactive and meaningful data visualization technique. Further- more, the report offers a method for a prediction of future events, which is demonstrated using the same visualization technique. The information embedded in the visualization

includes:

* The number of confirmed cases, deaths, mortality
* The change in time
* The countries affected
* Total vaccinations in different country
* Vaccinations based on age
* The proportionality between vaccinations and recovery

Our project aims to answer the question:

***How are the different countries with CoV-19 pandemic in relation to each other?***

**That is, which countries are in the danger zone and which are successfully managing the impacts?**

**The visualizations will be developed by roughly following the procedure as described by the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology.**

The phases include a business understanding, data understanding and preparation, modeling, evaluation, and deployment.

The visualization section is divided into two subsections.

Firstly, the visualization of the final analytical base table (ABT), which draws conclusions based on the current data.

And secondly the visualization of the predicted data.

Finally, the report will conclude with a short summary outlining the outcomes and recommendations of this report.

# Business Understanding

Many of the aspects defined in the CRISP-DM for the business understanding phase have been described in the previous section. **These include the background, objectives, and project plan. Other aspects, namely the requirements, constraints, and assumptions** are discussed in the following subsections.

## Requirements

As with any data analysis project, the most essential requirement is the actual data itself. The CoV-19 data should include all the variables under consideration as listed in the previous section. This data source should be reliable, and the content thereof verified. The data in this report will be processed by means of Python programming, Anaconda with Jupyter lab and Dash which is an open source statistical computing and graphing software. We use for data modelling for vaccination decision tree data mining algorithm

## Constraints

The data set time range is limited to the first official report in Wuhan, China on January 2020 up until the time that the modelling was completed on March 2024.

The countries [ Countries

Germany

USA

Italy

India

France

Spain

China

Algeria

Argentina

Continents

Europe

Asia

Africa

North America

South America

Oceania ] are limited to those which are contained in the data.

In other words, countries that are not listed in the data but have confirmed cases are not incorporated in this study. Some of the data is split by country and some by province. This makes comparisons easier since the proportions are roughly equal. For example, Germany is split up into provinces as this makes the data more comparable, for instance, to countries in Europe. However, some of the countries do not have provincial or state specific data.

## Assumptions

There are various assumptions regarding the scope of this investigation; these include:

* + 1. There is a difference between the number of confirmed cases in the data set and the actual number of people who are infected at a point in time.(The same goes for the number of deaths, and recoveries). The latter is most likely larger but the differences between the two are uncertain. This is due to various reasons, which include, amongst others, inadequate testing, self-isolation, and slow updates from medical facilities. This report assumes the number of confirmed cases as representative of the actual cases.
    2. There are random fluctuations or noise in the data regarding the number of cases, deaths, and recoveries, which do not necessarily have an underlying reason.
    3. Vaccination (Total vaccination based on country, types of vaccination,effect of vaccination, based on age, recovery rate)
    4. The countries affected ( test case due to covid )
    5. The data collected from the source is accurate (the reputability of the source is verified ).

# Data Understanding

The data understanding section consists of the collection, description, exploration, and

verification of the CoV-19 data(vaccination data).

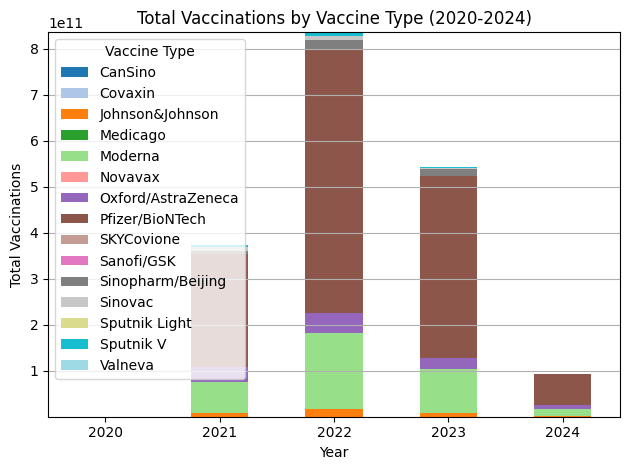
## Data Collection

The raw data for the CoV-19 pandemic is obtained from a GitHub data repository [ ] which is compiled and continuously updated by the Johns Hopkins University Center for Science and Engineering from various sources including the World Health Organisation (WHO). The data set is widely used on multiple reputable platforms such as the Humanitarian Data Exchange [[1]](#_bookmark36) and Kaggle [[2]](#_bookmark37) which verifies that the data is trusted.

Note that the values for the confirmed cases, deaths, and recoveries are cumulative and contained in separate CSV files.(The first four records for the CSV file of the confirmed cases are shown below in table [1.](#_bookmark7) The data sets will have to be integrated (discussed in section 4) to form the final ABT.

Table 1: Raw data for confirmed cases

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| country | year | affected | continent | percent\_affected | month |
| India | 2020 | 3214 | Asia | 0.034 | 1 |
| India | 2020 | 4856 | Asia | 0.051 | 2 |
| India | 2020 | 7123 | Asia | 0.075 | 3 |
| India | 2020 | 12589 | Asia | 0.027 | 4 |
| India | 2020 | 15482 | Asia | 0.058 | 5 |
| India | 2020 | 18791 | Asia | 0.093 | 6 |
| India | 2020 | 19547 | Asia | 0.132 | 7 |
| India | 2020 | 29873 | Asia | 0.104 | 8 |
| India | 2020 | 62218 | Asia | 0.066 | 9 |
| India | 2020 | 65871 | Asia | 0.041 | 10 |
| Csv (affected\_country.csv) |  |  |  |  |  |



Total vaccination vs year

## Description

The raw data is structured and contains ------------- records which comprise the data of -------1 different countries. The data is in fact, non-stationary and the source is updated daily. However, in this report, the data will be handled in a stationary manner. The variables and their corresponding data types contained in the raw data include:

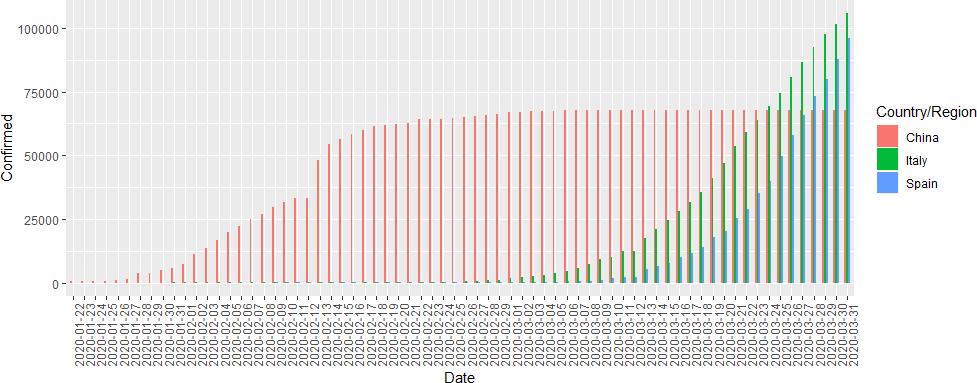
* Continent (nominal)
* Country (nominal)
* Date (interval)
* Confirmed Cases (numerical)
* Deaths (numerical)
* Mortality (numerical)
* Vaccinations(numerical)

The initial ABT will also include a continent variable for visualization purposes. The continent for each country is added in the ABT containing country names .Date are group by in to year.

## Exploratory Data Analysis (EDA)

The confirmed cases, deaths, and vaccinations are most significant in terms of exploratory data analysis. There is a general relationship between the numerical variables but these vary for different countries. For example, the data increases as time passes and the deaths or recoveries are always lower than the confirmed cases. But some countries have more deaths than recoveries at a point in time and vice versa.

To further illustrate this, the confirmed cases, deaths and recoveries are plotted over a time series for three different countries or provinces namely China , Italy, and Spain ,Germany, India ,Algeria ,France . These countries are specifically chosen for the EDA illustrations because they have the highest value range which provides a sense of the data limits. Furthermore, they are also roughly representative of the distribution pattern in the rest of the data.



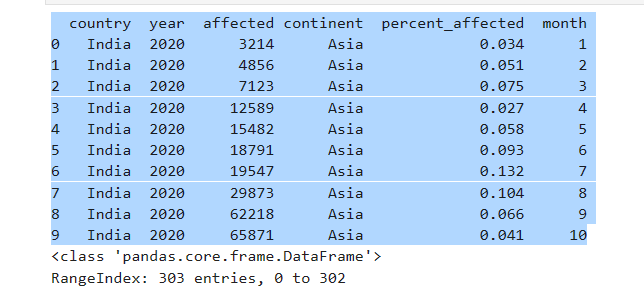
The raw data also includes some cases were only a few days were recorded and the

corresponding confirmed cases, deaths, are very low or zero. These are discarded because they have no qualitative purpose for the final visualizations.

Data Transformation

Data transformations includes cleaning, constructing, integrating and formatting the data selected. The data contained in the initial ABT is already clean in the sense that there are no quality issues or any missing values (NA values for province or state means that the country is considered exclusively).

Aggregated (group by function is used in vaccinations .csv ) to extract year and analysis is based on year. Total case, vaccination rate based on age group, comparison of vaccination are done based on aggregated function.



The data has been integrated by combining the confirmed cases, vaccination and by assigning continents to each country .

# Modelling

This section focuses on using the final ABT data to develop a model to make predictions. More specifically, the model is used to predict the confirmed cases, deaths, and vaccination .

Predicting the outcomes of a pandemic is extremely complicated and requires complex detailed simulations . The model in this study will be trained based on the current data. However, the occurrence of specific events after the time range which the data was trained on can have a significant impact on the data trend. For example, the decision by a government to implement a regional or countrywide quarantine will have a significant impact on the new confirmed cases. If the model is trained without the circumstances of a quarantine it will produce poor and over inflated predictions. Another example, the vaccination rates are highly dependent on health system efficiency and treatment capacity. **The death rate could be low because the confirmed cases are below treatment capacity. However, when the confirmed cases exceed the treatment capacity, the death rate should increase drastically.**

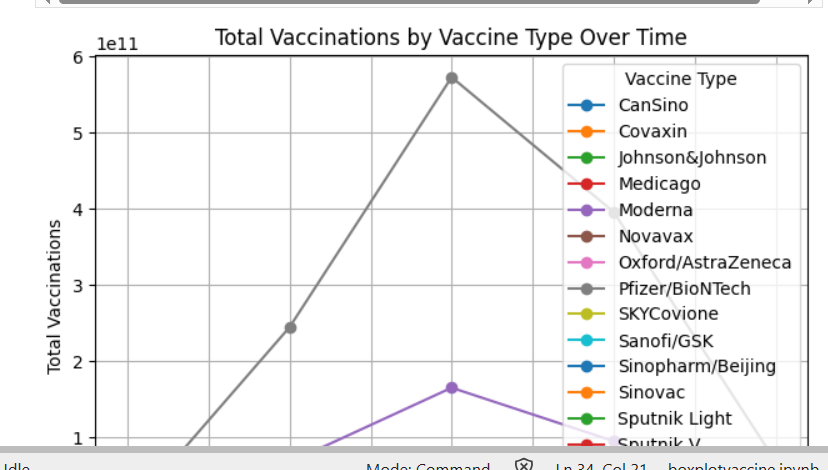
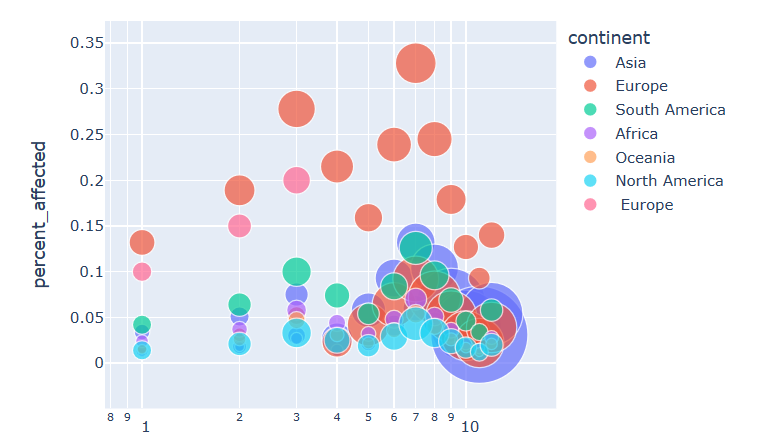


Fig 2 vaccination in different year

# Evaluation

# Similar to that of the modelling section, this section will only discuss the predictions for the confirmed cases and vaccination.Fig. [2](#_bookmark26) shows a plot of the predicted vaccination cases for each country under investigation and for its consecutive 4 years.

# 



# Deployment

The deployment phase of the project life cycle involves presenting the information gathered from the data analysis and modelling phase. For this report, the deployment phase includes the development of a dashboard.

# Conclusion

This report has clearly defined a topic and the objectives were addressed appropriately. The visualisations were developed by means of the CRISP-DM process and relevance of each phase has been accurately described. The visualisations presented in this report can be utilised to answer case-specific queries or it can be used to gain a broad insight into the pandemic. The predictions made by this report are fairly accurate, however, they are only generated for three days past the most recent date in the data set. The models are therefore not recommended for beyond these limits.