 **UNIVERSITY CARLOS III OF MADRID**

**ARISTOTLE UNIVERSITY OF THESSALONIKI**

**FACULTY OF SCIENCES**

**DEPARTMENT OF MATHEMATICS**

**INTERNSHIP**

**DATA MINING IN THE DATASETS**

**OF COVID-19**

**ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΣΕ ΣΕΤ ΔΕΔΟΜΕΝΩΝ**

**ΤΟΥ COVID-19**

**ALEXOUDI PANAGIOTA**

**AEM: 15667**

**SUPERVISOR:**

**JESUS GARCIA, PROFESSOR**

**JUAN PEDRO LLERENA CAÑA, PHD CANDIDATE**

**MADRID 2020**

# Abstract

In this internship

# Περίληψη

Στη παρούσα πρακτική

**Table of Contents**

[Abstract 3](#_Toc47547938)

[Περίληψη 5](#_Toc47547939)

[Acknowledgements 9](#_Toc47547940)

[Chapter 1: Introduction 11](#_Toc47547941)

[Chapter 2: Introduction - History of Data Mining 12](#_Toc47547942)

[2.1 Data mining 12](#_Toc47547943)

[2.2 Roots of Data Mining 13](#_Toc47547944)

[2.3 Steps of processing 14](#_Toc47547945)

[2.4 Data mining techniques 16](#_Toc47547946)

[2.5 Challenges of Data Mining 17](#_Toc47547947)

[Chapter 3: Coronavirus Disease 2019 19](#_Toc47547948)

[3.1 Definition of Covid-19 19](#_Toc47547949)

[3.2 History of Covid-19 21](#_Toc47547950)

[3.2.1 COVID-19 pandemic by country and territory and cases 22](#_Toc47547951)

[3.2.2 COVID-19 pandemic deaths and Mortality due to COVID-19 23](#_Toc47547952)

[3.2.3 Gendered impact of the COVID-19 pandemic (Sex differences) 24](#_Toc47547953)

[3.3 Symptoms of Covid-19 25](#_Toc47547954)

[3.4 Transmission of Covid-19 26](#_Toc47547955)

[3.5 Methods of Diagnosis 27](#_Toc47547956)

[3.6 Cure and measures 27](#_Toc47547957)

[3.7 Consequences of Covid-19 pandemic 29](#_Toc47547958)

[Chapter 4: Application 30](#_Toc47547959)

[4.1 Visualization of datasets 31](#_Toc47547960)

[4.1.1 Heat map 31](#_Toc47547961)

[4.1.2 Choropleth maps 32](#_Toc47547962)

[4.1.3 Bar charts 34](#_Toc47547963)

[4.1.4 Tables and Choropleth map 36](#_Toc47547964)

[4.1.5 Graphs related with temperature (change and add) 37](#_Toc47547965)

[4.1.6 Relationship between temperature and confirmed cases 40](#_Toc47547966)

[4.2 Implementation of Data mining techniques 41](#_Toc47547967)

[4.2.1 Correlation 41](#_Toc47547968)

[4.2.2 Time series forecasting 41](#_Toc47547969)

[Chapter 5: Conclusion – Future Research 50](#_Toc47547970)

[Bibliography 51](#_Toc47547971)

# Acknowledgements

I would like to thank professor Jesus Garcia for the opportunity he gave me to work on this internship, as I have learned a lot regarding Data Mining. I would also like to thank PhD candidate Juan Pedro Llerena Caña for the collaboration and his help throughout this internship.

Finally, I would like to thank my parents, Vasileios and Lamprini for their support in every step of my life and the friends I met as the experience of the internship would be totally different without them.

# Chapter 1: Introduction

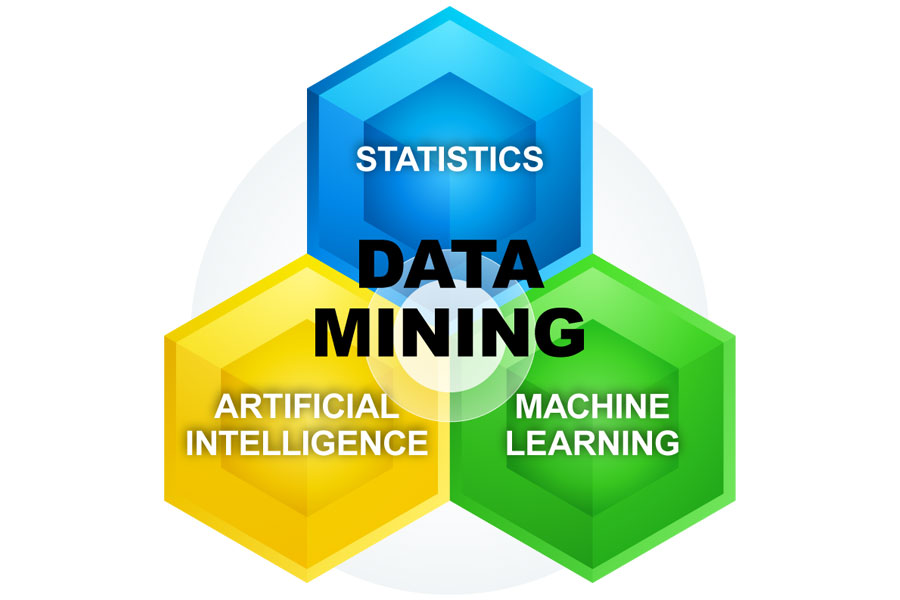
# Chapter 2: Introduction - History of Data Mining

## 2.1 Data mining

The computerization of our society has substantially enhanced our capabilities for both generating and collecting data from diverse sources(Larose & Larose, 2014). A tremendous amount of data has flooded almost every aspect of our lives. This explosive growth in stored or transient data has generated an urgent need for new techniques and automated tools that can intelligently assist us in transforming the vast amounts of data into useful information and knowledge. This has led to the generation of a promising and flourishing frontier in computer science called data mining and its various applications.

Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques.

Its foundation comprises three intertwined scientific disciplines: statistics (the numeric study of data relationships), artificial intelligence (human-like intelligence displayed by software and/or machines) and machine learning (algorithms that can learn from data to make predictions) (Figure 1). What was old is new again, as data mining technology keeps evolving to keep pace with the limitless potential of big data and affordable computing power. Data mining is widely used in business (insurance, banking, retail), science research (astronomy, medicine), and government security (detection of criminals and terrorists). One of the earliest successful applications of data mining was credit-card-fraud detection.

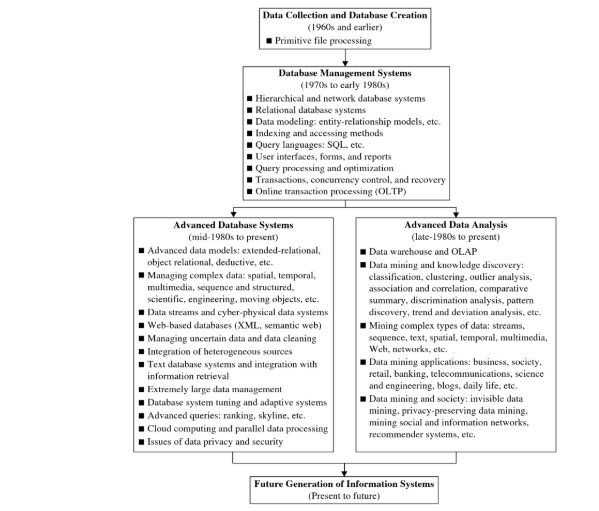


*Figure 1: Scientific disciplines that comprise the foundations of data mining*

## 2.2 Roots of Data Mining

Data mining can be viewed as the result of the natural evolution of information technology. The database and data management industry evolved in the development of several critical functionalities, data collection and data base creation, data management (including data storage and retrieval and database transaction processing) and advanced data analysis (involving data warehousing and data mining) (Figure 2).

As computer storage capacities increased during the 1980s, many companies began to store more transactional data. The resulting record collections, often called data warehouses, were too large to be analyzed with traditional statistical approaches. Several computer science conferences and workshops were held to consider how recent advances in the field of artificial intelligence (AI)—such as discoveries from expert systems, genetic algorithms, machine learning, and neural networks—could be adapted for knowledge discovery (the preferred term in the computer science community). The process led in 1995 to the First International Conference on Knowledge Discovery and Data Mining, held in Montreal, and the launch in 1997 of the journal Data Mining and Knowledge Discovery. This was also the period when many early data-mining companies were formed and products were introduced. Nowadays numerous database systems offer query and transaction processing as common practice. Advanced data analysis has naturally become the next step.



*Figure 2: The evolution of database system technology*

## 2.3 Steps of processing

The complete data-mining process involves multiple steps, from understanding the goals of a project and what data are available to implementing process changes based on the final analysis.

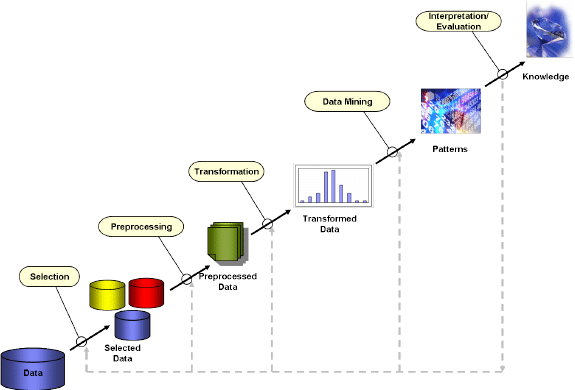
The knowledge discovery in databases (KDD) process is commonly defined with the following stages (Figure 3):

1. Data selection (where data relevant to the analysis are retrieved from the database)
2. Data preprocessing:

a. Data cleaning (to remove noise and inconsistent data)

b. Data integration (where multiple data sources may be combined)

1. Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)
2. Data mining techniques (an essential process where intelligent methods are applied to extract data patters)
3. Pattern evaluation (to identify the truly interesting patterns representing knowledge based or interestingness measures)
4. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)



*Figure 3: Data mining process steps*

## 2.4 Data mining techniques

Data mining involves seven common classes of techniques (Figure 4)(Han, Pei, & Kamber, 2011):

1.Classification: This analysis is used to retrieve important and relevant information about data, and metadata. This data mining method helps to classify data in different classes. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

2. Clustering: Clustering analysis is a data mining technique to identify groups and structures in the data that are in some way or another "similar", without using known structures in the data. This process helps to understand the differences and similarities between the data.

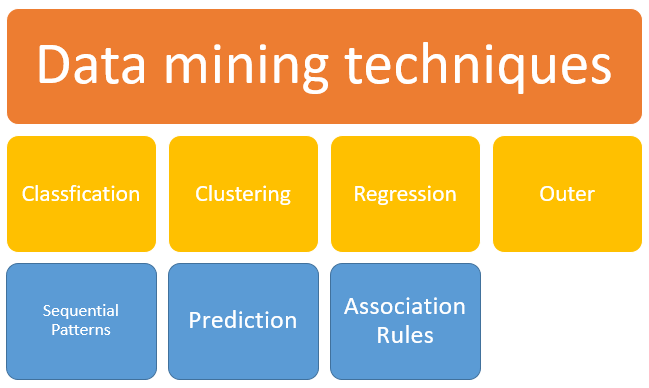
3. Regression: Regression analysis is the data mining method of identifying and analyzing the relationship between variables. It is used to identify the likelihood of a specific variable, given the presence of other variables.

4. Association Rules: This data mining technique helps to find the association between two or more Items. It discovers a hidden pattern in the data set.

5. Outer/Anomaly detection: This type of data mining technique refers to observation of data items in the dataset which do not match an expected pattern or expected behavior. This technique can be used in a variety of domains, such as intrusion, detection, fraud or fault detection, etc. Outer detection is also called Outlier Analysis or Outlier mining.

6. Sequential Patterns: This data mining technique helps to discover or identify similar patterns or trends in transaction data for certain period.

7. Prediction: Prediction has used a combination of the other data mining techniques like trends, sequential patterns, clustering, classification, etc. It analyzes past events or instances in a right sequence for predicting a future event.



*Figure 4: Data mining techniques*

## 2.5 Challenges of Data Mining

While a powerful process, data mining is hindered by the increasing quantity and complexity of big data. Where data are collected by firms every day, decision-makers need ways to extract, analyze, and gain insight from their abundant repository of data.

1. Big Data: The challenges of big data are prolific and penetrate every field that collects, stores, and analyzes data. Big data is characterized by four major challenges: volume, variety, veracity, and velocity. The goal of data mining is to mediate these challenges and unlock the data’s value.

2. Over-Fitting Models: Over-fitting occurs when a model explains the natural errors within the sample instead of the underlying trends of the population. Over-fitted models are often overly complex and utilize an excess of independent variables to generate a prediction. Therefore, the risk of over-fitting is heighted by the increase in volume and variety of data. Too few variables make the model irrelevant, where as too many variables restrict the model to the known sample data. The challenge is to moderate the number of variables used in data mining models and balance its predictive power with accuracy.

3. Cost of Scale: As data velocity continues to increase data’s volume and variety, firms must scale these models and apply them across the entire organization. Unlocking the full benefits of data mining with these models requires significant investment in computing infrastructure and processing power. To reach scale, organizations must purchase and maintain powerful computers, servers, and software designed to handle the firm’s large quantity and variety of data.

4. Privacy and Security: The increased storage requirement of data has forced many firms to turn toward cloud computing and storage. While the cloud has empowered many modern advances in data mining, the nature of the service creates significant privacy and security threats. Organizations must protect their data from malicious figures to maintain the trust of their partners and customers.

With data privacy comes the need for organizations to develop internal rules and constraints on the use and implementation of a customer’s data. Data mining is a powerful tool that provides businesses with compelling insights into their consumers. Organizations must weigh this relationship with their customers, develop policies to benefit consumers, and communicate these policies to the consumers to maintain a trustworthy relationship.

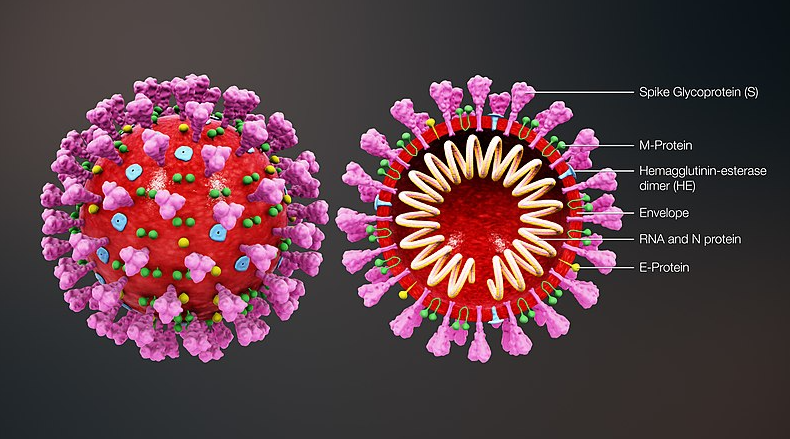
# Chapter 3: Coronavirus Disease 2019

## 3.1 Definition of Covid-19

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). SARS-CoV-2 is the third coronavirus, after SARS-CoV1 and MERS-CoV.

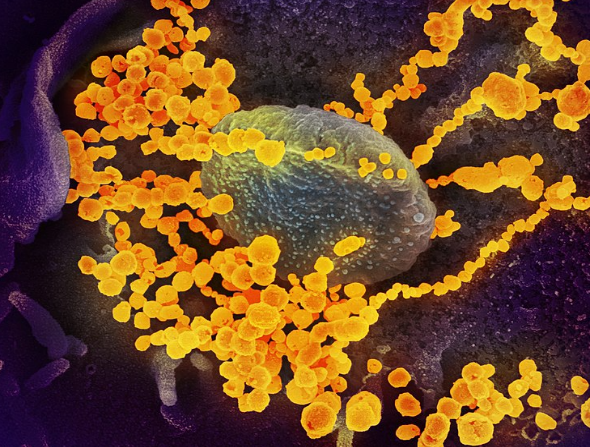
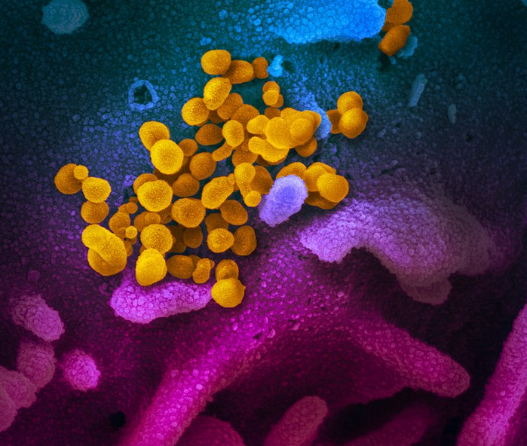
It is believed to have zoonotic origins and has close genetic similarity to bat coronaviruses, suggesting it emerged from a bat-borne virus. There is no evidence yet to link an intermediate animal reservoir, such as a pangolin, to its introduction to humans. The virus shows little genetic diversity, indicating that the spillover event introducing SARS-CoV-2 to humans is likely to have occurred in late 2019.

As described by the U.S. National Institutes of Health, it is the successor to SARS-CoV-1. SARS-CoV-2 is a positive-sense single-stranded RNA virus.



*Figure 5: Structural view of a coronavirus*

Each SARS-CoV-2 virion is 50–200 nanometres in diameter. Like other coronaviruses, SARS-CoV-2 has four structural proteins, known as the S (spike), E (envelope), M (membrane), and N (nucleocapsid) proteins; the N protein holds the RNA genome, and the S, E, and M proteins together create the viral envelope (Figure 5). The spike protein, which has been imaged at the atomic level using cryogenic electron microscopy, is the protein responsible for allowing the virus to attach to and fuse with the membrane of a host cell; specifically, its S1 subunit catalyzes attachment, the S2 subunit fusion (Figure 6).

*Figure 6: Digitally colourised scanning electron micrographs of SARS-CoV-2 virions (yellow) emerging from human cells cultured in a laboratory*

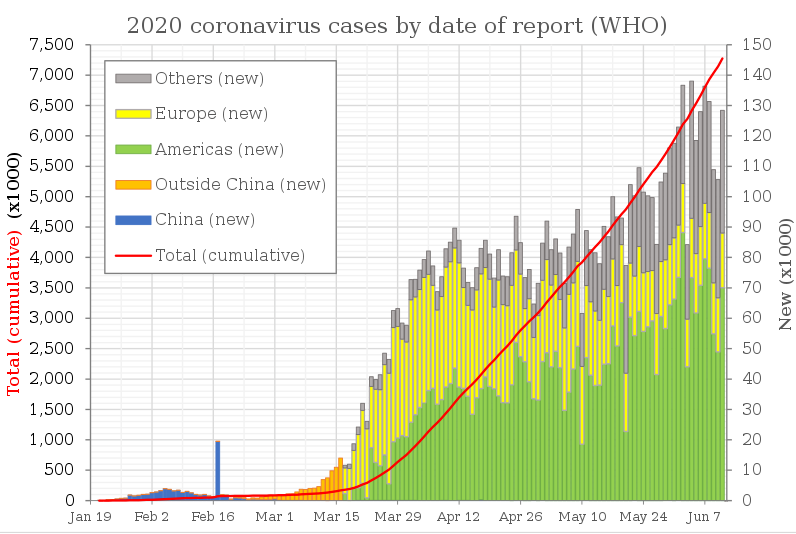
Protein modeling experiments on the spike protein of the virus soon suggested that SARS-CoV-2 has sufficient affinity to the receptor angiotensin converting enzyme 2 (ACE2) on human cells to use them as a mechanism of cell entry. By 22 January 2020, a group in China working with the full virus genome and a group in the United States using reverse genetics methods independently and experimentally demonstrated that ACE2 could act as the receptor for SARS-CoV-2. Studies have shown that SARS-CoV-2 has a higher affinity to human ACE2 than the original SARS virus strain SARS-CoV-2 may also use basigin to assist in cell entry.

Initial spike protein priming by transmembrane protease, serine 2 (TMPRSS2) is essential for entry of SARS-CoV-2. After a SARS-CoV-2 virion attaches to a target cell, the cell's protease TMPRSS2 cuts open the spike protein of the virus, exposing a fusion peptide in the S2 subunit, and the host receptor ACE2. After fusion, an endosome forms around the virion, separating it from the rest of the host cell. The virion escapes when the pH of the endosome drops or when cathepsin, a host cysteine protease, cleaves it. The virion then releases RNA into the cell and forces the cell to produce and disseminate copies of the virus, which infect more cells.

## 3.2 History of Covid-19

The COVID-19 pandemic, also known as the coronavirus pandemic, is an ongoing pandemic of coronavirus disease 2019. Based on the low variability exhibited among known SARS-CoV-2 genomic sequences, the strain is thought to have been detected by health authorities within weeks of its emergence among the human population in late 2019.

The outbreak was first identified in Wuhan, China, in December 2019 and has resulted in an ongoing pandemic. The first case may be traced back to 17 November 2019(Lu, Stratton, & Tang, 2020). The World Health Organization declared the outbreak a Public Health Emergency of International Concern on 30 January, and a pandemic on 11 March. Local transmission of the disease has occurred in most countries across all six WHO regions. The virus subsequently spread to all provinces of China and to more than 150 other countries in Asia, Europe, North America, South America, Africa, and Oceania. Human-to-human transmission of the virus has been confirmed in all these regions. As of 12 June 2020, more than 7.5 million cases of COVID-19 have been reported in more than 188 countries and territories, resulting in more than 421,000 deaths; more than 3.53 million people have recovered (Figure 7).



*Figure 7: Epidemic curve of COVID-19 by date of report*

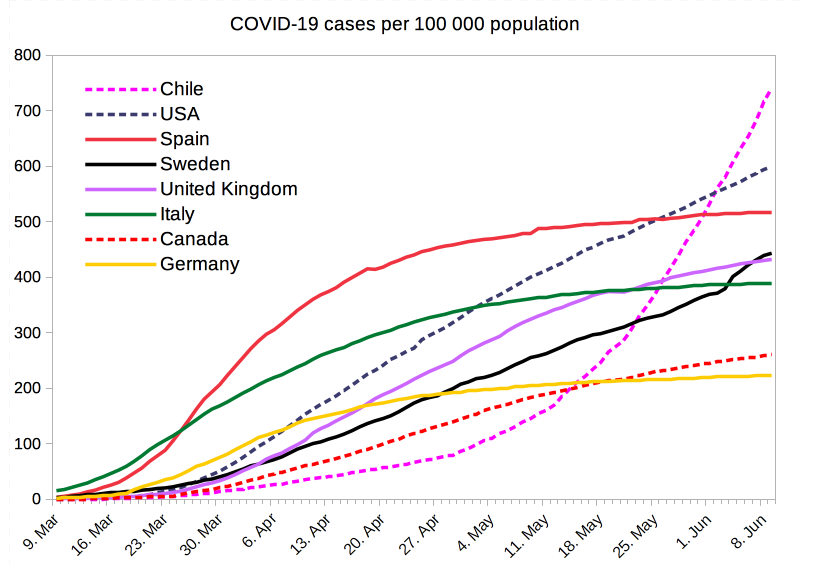
### 3.2.1 COVID-19 pandemic by country and territory and cases

Cases refer to the number of people who have been tested for COVID-19, and whose test have been confirmed positive according to official protocols. As of 24 May, countries that published their testing data have typically performed many tests equal to 2.6 percent of their population, while no country has tested samples equal to more than 17.3 percent of its population. Many countries, early on, had official policies to not test those with only mild symptoms. An analysis of the early phase of the outbreak up to 23 January estimated 86 percent of COVID-19 infections had not been detected, and that these undocumented infections were the source for 79 percent of documented cases. Several other studies, using a variety of methods, have estimated that numbers of infections in many countries are likely to be considerably greater than the reported cases.

On 9 April 2020, preliminary results found that 15 percent of people tested in Gangelt, the center of a major infection cluster in Germany, tested positive for antibodies. Screening for COVID-19 in pregnant women in New York City, and blood donors in the Netherlands, has also found rates of positive antibody tests that may indicate more infections than reported. However, such antibody surveys can be unreliable due to a selection bias in who volunteers to take the tests, and due to false positives. Some results (such as the Gangelt study) have received substantial press coverage without first passing through peer review.

Analysis by age in China indicates that a relatively low proportion of cases occur in individuals under 20. It is not clear whether this is because young people are less likely to be infected, or less likely to develop serious symptoms and seek medical attention and be tested. A retrospective cohort study in China found that children were as likely to be infected as adults. Countries that test more, relative to the number of deaths, have a younger age distribution of cases, relative to the wider population.

Initial estimates of the basic reproduction number (R0) for COVID-19 in January were between 1.4 and 2.5, but a subsequent analysis has concluded that it may be about 5.7 (with a 95 percent confidence interval of 3.8 to 8.9). R0 can vary across populations and is not to be confused with the effective reproduction number (commonly just called R), which takes into account effects such as social distancing and herd immunity. As of mid-May 2020, the effective R is close to or below 1.0 in many countries, meaning the spread of the disease in these areas is stable or decreasing.



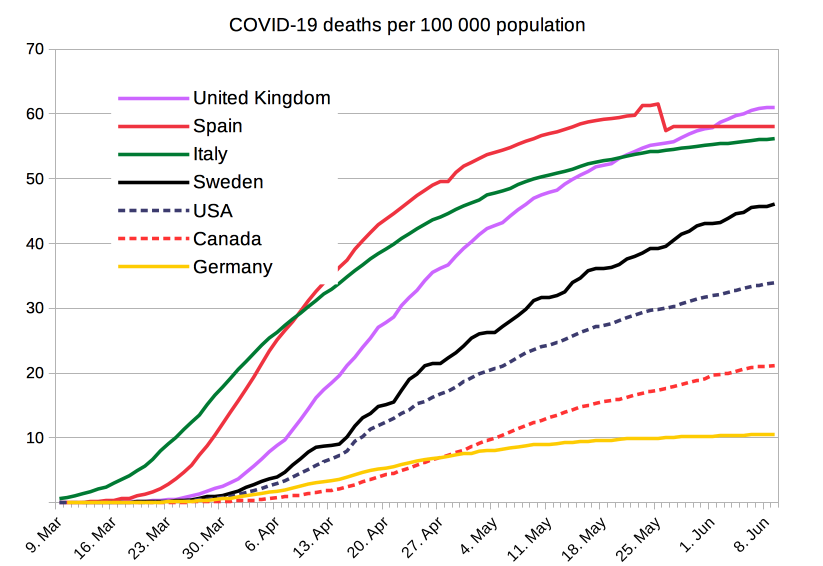
*Figure 8: COVID-19 total cases per 100 000 populations from selected countries*

### 3.2.2 COVID-19 pandemic deaths and Mortality due to COVID-19

Most people who contract COVID-19 recover. For those who do not, the time between the onset of symptoms and death usually ranges from 6 to 41 days, typically about 14 days. As of 12 June 2020, approximately 421,000 deaths had been attributed to COVID-19. In China, as of 5 February, about 80 percent of deaths were recorded in those aged over 60, and 75 percent had pre-existing health conditions including cardiovascular diseases and diabetes.

The first confirmed death was in Wuhan on 9 January 2020. The first death outside of China occurred on 1 February in the Philippines, and the first death outside Asia was in France on 14 February.

Official deaths from COVID-19 generally refer to people who died after testing positive according to protocols. This may ignore deaths of people who die without testing, e.g. at home or in nursing homes. Conversely, deaths of people who had underlying conditions may lead to over counting. Comparison of statistics for deaths for all causes versus the seasonal average indicates excess mortality in many countries. In the worst affected areas, mortality has been several times higher than average. In New York City, deaths have been four times higher than average, in Paris twice as high, and in many European countries, deaths have been on average 20 to 30 percent higher than normal. This excess mortality may include deaths due to strained healthcare systems and bans on elective surgery.



*Figure 9: COVID-19 deaths per 100 000 populations from selected countries*

### 3.2.3 Gendered impact of the COVID-19 pandemic (Sex differences)

The impact of the pandemic and its mortality rate are different for men and women. Mortality is higher in men in studies conducted in China and Italy. The higher risk for men appears in their 50s, and begins to taper off only at 90. In China, the death rate was 2.8 percent for men and 1.7 percent for women The exact reasons for this sex-difference are not known, but genetic and behavior factors could be a reason. Sex-based immunological differences, a lower prevalence of smoking in women, and men developing co-morbid conditions such as hypertension at a younger age than women could have contributed to the higher mortality in men.

In Europe, of those infected with COVID‑19, 57% were men; of those infected with COVID‑19 who also died, 72% were men. As of April 2020, the U.S. government is not tracking sex-related data of COVID‑19 infections. Research has shown that viral illnesses like Ebola, HIV, influenza, and SARS affect men and women differently. A higher percentage of health workers, particularly nurses, are women, and they have a higher chance of being exposed to the virus. School closures, lockdowns, and reduced access to healthcare following the COVID-19 pandemic may differentially affect the genders and possibly exaggerate existing gender disparity.

## 3.3 Symptoms of Covid-19

Coronavirus disease is an infectious disease caused by a newly discovered coronavirus, therefore the symptoms vary in each case.

Fever is the most common symptom of COVID-19, but is highly variable in severity and presentation, with some older, immunocompromised, or critically ill people not having fever at all.

Other common symptoms include cough, loss of appetite, fatigue, shortness of breath, sputum production, and muscle and joint pains. Symptoms such as nausea, vomiting, and diarrhea have been observed in varying percentages. Less common symptoms include sneezing, runny nose, sore throat, and skin lesions. Some cases in China initially presented with only chest tightness and palpitations. A decreased sense of smell or disturbances in taste may occur. Loss of smell was a presenting symptom in 30% of confirmed cases in South Korea.

Among those who develop symptoms, approximately one in five may become more seriously ill and have difficulty breathing. Emergency symptoms include difficulty breathing, persistent chest pain or pressure, sudden confusion, difficulty waking, and bluish face or lips; immediate medical attention is advised if these symptoms are present(Wang et al., 2020). Further development of the disease can lead to complications including pneumonia, acute respiratory distress syndrome, sepsis, septic shock, and kidney failure(Chen et al., 2020).

As is common with infections, there is a delay between the moment a person is first infected and the time he or she develops symptoms. This is called the incubation period. The typical incubation period for COVID‑19 is five or six days, but it can range from one to fourteen days with approximately ten percent of cases taking longer.

Most people infected with the COVID-19 virus will experience mild to moderate respiratory illness and recover without requiring special treatment. Older people, and those with underlying medical problems like cardiovascular disease, diabetes, chronic respiratory disease, and cancer are more likely to develop serious illness.

Some infected people have no symptoms, known as asymptomatic or presymptomatic carriers; transmission from such a carrier is considered possible. As at 6 April, estimates of the asymptomatic ratio range widely from 5% to 80%.

## 3.4 Transmission of Covid-19

Human-to-human transmission of SARS-CoV-2 was confirmed on 20 January 2020, during the COVID-19 pandemic. Transmission occurs primarily via respiratory droplets from coughs and sneezes within a range of about 1.8 meters (6 ft). The droplets usually fall to the ground or onto surfaces rather than travelling through air over long distances. Epidemiological studies estimate each infection results in 1.4 to 3.9 new ones when no members of the community are immune and no preventive measures taken.

Less commonly, people may become infected by touching a contaminated surface and then touching their face. Preliminary research indicates that the virus may remain viable on plastic (polypropylene) and stainless steel (AISI 304) for up to three days, but does not survive on cardboard for more than one day or on copper for more than four hours; the virus is inactivated by soap, which destabilizes its lipid bilayer. Viral RNA has also been found in stool samples and semen from infected individuals. It is most contagious during the first three days after the onset of symptoms, although spread is possible before symptoms appear, and from people who do not show symptoms(Li et al., 2020; Rothe et al., 2020).

There is some evidence of human-to-animal transmission of SARS-CoV-2, including examples in felids. Some institutions have advised those infected with SARS-CoV-2 to restrict contact with animals.

## 3.5 Methods of Diagnosis

The WHO has published several testing protocols for the disease. The standard method of diagnosis is real-time reverse transcription polymerase chain reaction (rRT-PCR) from a nasopharyngeal swab. The test is typically done on respiratory samples obtained by a nasopharyngeal swab; however, a nasal swab or sputum sample may also be used. Results are generally available within a few hours to two days.

Blood tests can be used, but these require two blood samples taken two weeks apart, and the results have little immediate value. Chinese scientists were able to isolate a strain of the coronavirus and publish the genetic sequence so laboratories across the world could independently develop polymerase chain reaction (PCR) tests to detect infection by the virus. As of 4 April 2020, antibody tests (which may detect active infections and whether a person had been infected in the past) were in development, but not yet widely used. The Chinese experience with testing has shown the accuracy is only 60 to 70%. The FDA in the United States approved the first point-of-care test on 21 March 2020 for use at the end of that month.

Along with laboratory testing, chest CT scans may be helpful to diagnose COVID‑19 in individuals with a high clinical suspicion of infection but are not recommended for routine screening. Bilateral multilobar ground-glass opacities with a peripheral, asymmetric, and posterior distribution are common in early infection. Subpleural dominance, crazy paving (lobular septal thickening with variable alveolar filling), and consolidation may appear as the disease progresses.

## 3.6 Cure and measures

According to the World Health Organization (WHO), there are no vaccines nor specific antiviral treatments for COVID-19. Therefore, management involves the treatment of symptoms, supportive care, isolation, and experimental measures.

Recommended measures to prevent infection include frequent hand washing, maintaining physical distance from others (especially from those with symptoms), quarantine (especially for those with symptoms), covering coughs, washing hands with soap and water often and for at least 20 seconds, practicing good respiratory hygiene, and avoiding touching the eyes, nose, or mouth with unwashed hands (Figure 10). Outside the human body, the virus is killed by household soap, which bursts its protective bubble. The use of cloth face coverings such as a scarf or a bandana is recommended in public settings to minimize the risk of transmissions, with some authorities requiring their use. Medical grade facemasks such as N95 masks should only be used by healthcare workers, first responders and those who care for infected individuals.



*Figure 10: Measures to prevent transmission of the virus*

Social distancing strategies aim to reduce contact of infected persons with large groups by closing schools and workplaces, restricting travel, and cancelling large public gatherings. Distancing guidelines also include that people stay at least 6 feet (1.8 m) apart. After the implementation of social distancing and stay-at-home orders, many regions have been able to sustain an effective transmission rate ("Rt") of less than one, meaning the disease is in remission in those areas.

Authorities worldwide have responded by implementing travel restrictions, lockdowns, workplace hazard controls, and facility closures. Many places have also worked to increase testing capacity and trace contacts of infected persons.

As a COVID-19 vaccine is not expected until 2021 at the earliest, a key part of managing COVID‑19 is trying to decrease and delay the epidemic peak, known as "flattening the curve". This is done by slowing the infection rate to decrease the risk of health services being overwhelmed, allowing for better treatment of current cases, and delaying additional cases until effective treatments or a vaccine become available.

At this time, there are no specific vaccines or treatments for COVID-19. However, there are many ongoing clinical trials evaluating potential treatments. WHO will continue to provide updated information as soon as clinical findings become available.

## 3.7 Consequences of Covid-19 pandemic

The pandemic has caused immense global social and economic disruption, including the largest global recession since the Great Depression. It has led to the postponement or cancellation of sporting, religious, political, and cultural events, widespread supply shortages exacerbated by panic buying, and decreased emissions of pollutants and greenhouse gases. Schools, universities, and colleges have been closed either on a nationwide or local basis in 172 countries, affecting approximately 98.5 percent of the world's student population. Misinformation about the virus has circulated through social media and the mass media. There have been incidents of xenophobia and discrimination against Chinese people and against those perceived as being Chinese or as being from areas with high infection rates

# Chapter 4: Application

During this internship we worked on Data Mining on the datasets of the Covid-19. The main idea was to implement data mining techniques on the available datasets and extract information and conclusions from them. The datasets come from various sources like the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, the World Bank and WHO including the use of Kaggle.

The tools we primarily used for this project are the Jupyter Notebook, which is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and explanatory text. For this purpose, the programming language we used is Python and we imported many different packages and libraries, for example numpy and scipy.

The process followed to achieve the outcome is divided into the following steps:

1. Selection/Target Data
2. Pre-processing
3. Transformation
4. Implement Data mining techniques
5. Interpretation/Evaluation

The first step is really important as there is a vast amount of datasets online with infinite information. After finding all the necessary datasets we had to set specific goals and the possible outcomes we would like to explore. Therefore, we represented these datasets using visualization tools to identify patterns. Pre-processing the data is also significant. In this way we can detect any outliers and remove noise as well as to detect any missing values. The next step involves Transformation. In this step we accumulated cases during 15 days after getting 1 cases per 100.000 inhabitants as a standardization. (change)

Prioritize the objectives and find possible techniques to implement is the next step. In this project we used Correlation and Prediction. In particular, we wanted to identify the correlation between the number of cases, deaths and different meteorological variables like Temperature and Fog presence. In Prediction we wanted to find a time series model to forecast future cases. (write more/change it)

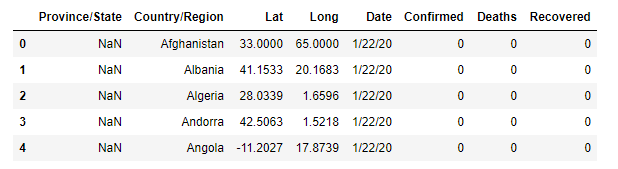
The full code is uploaded on my GitHub account:

<https://github.com/PanagiotaAlex/Covid-19-project/tree/master>

Many of the graphs are interactive so you can check them there. In case some of the plots are not presented you can copy the link and paste it here: <https://nbviewer.jupyter.org/>

## 4.1 Visualization of datasets

The first dataset we used contains information about the Confirmed cases, the number of Deaths and Recoveries according to Country/Region and Date. The Latitude and Longitude of each Country is given as well as the Province/State which we are not going to use as the values throughout that column are NaN (Figure 11).

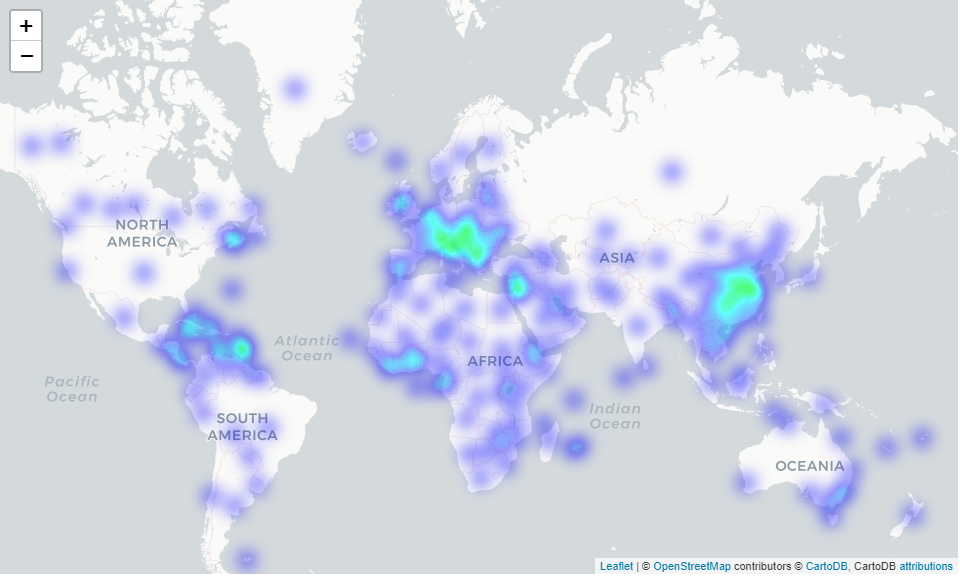


*Figure 11: Dataset representation*

We want to check the spread of COVID-19 geospatially by plotting running maps to observe the spread of confirmed cases and deaths as well as the recovered cases.

### 4.1.1 Heat map

We created a basic heat map using the Latitude and Longitude of each country. For this reason, we used Folium. Folium is a Python Library that can allow us to visualize spatial data in an interactive manner, straight within the notebooks environment. In Figure 12 we can see the initial spread of Coronavirus all over the world. With vivid green color are the places where the number of cases are higher, for instance China and the south Europe like Italy and Spain.

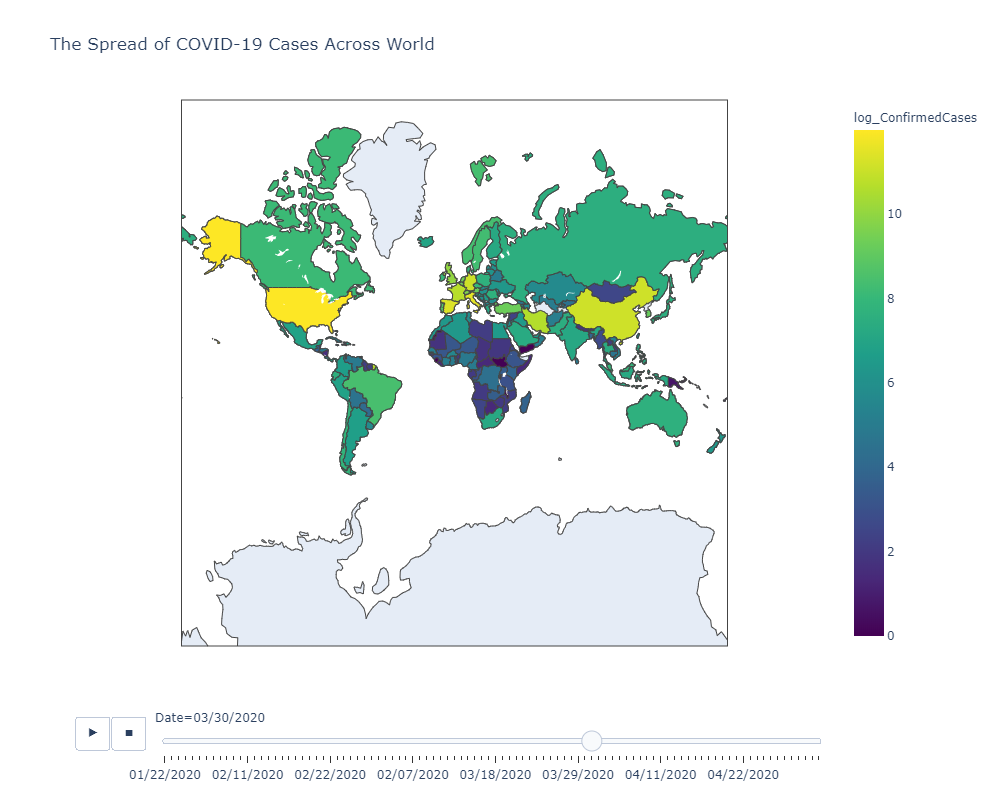


*Figure 12: Heat map of the world*

### 4.1.2 Choropleth maps

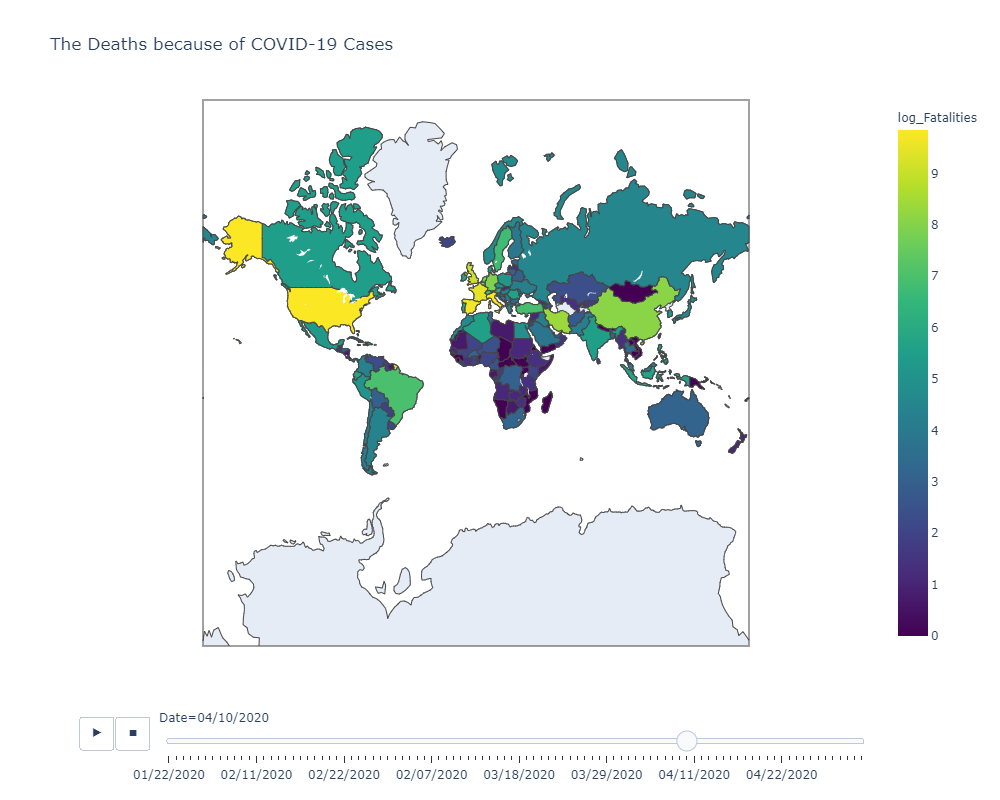
In order to have a better understanding of our data we started processing them. First of all, we grouped the same countries with their successive dates together and replaced NaN values in Province/State as Not Reported. The dates are from 22/01/2020 until 26/04/2020. Then we plotted an interactive map to observe the spread of confirmed cases and a map to observe the spread of deaths in a log scale. In order to achieve this, we grouped the number of confirmed cases and deaths according to date. As a visualization technique we used a choropleth map to plot the figure. Choropleth Maps display divided geographical areas or regions that are colored, shaded or patterned in relation to a data variable.

In Figure 13 we can see the spread of the virus around the world. The yellow color indicates the highest numbers in scale while dark blue the lowest numbers. At the end of January China is the most affected country from the virus. In the beginning of March, the virus spreads across other countries as well, with Europe and the USA reporting more and more cases. By the end of March, the virus has spread across all continents with Africa being the least affected. While the number of cases in Chine drop significantly the same numbers rise in the USA. In April the top affected country is the USA while the spread in Europe remains in the same levels or slightly rise.



*Figure 13: Spread of confirmed case across the world*

In Figure 14 we see the spread of deaths around the world. The difference between the spread of cases and deaths is significant as the fatality rate is lower. Initially China is the only country with reported deaths at the end of January. By the end of February, we have the first deaths in Italy in particular. In the middle of March countries like Spain, Iran and the USA begin to have high rates of fatality. In the middle of April, the number of deaths decrease in China and rise in Spain, Italy, France, the UK and the USA.

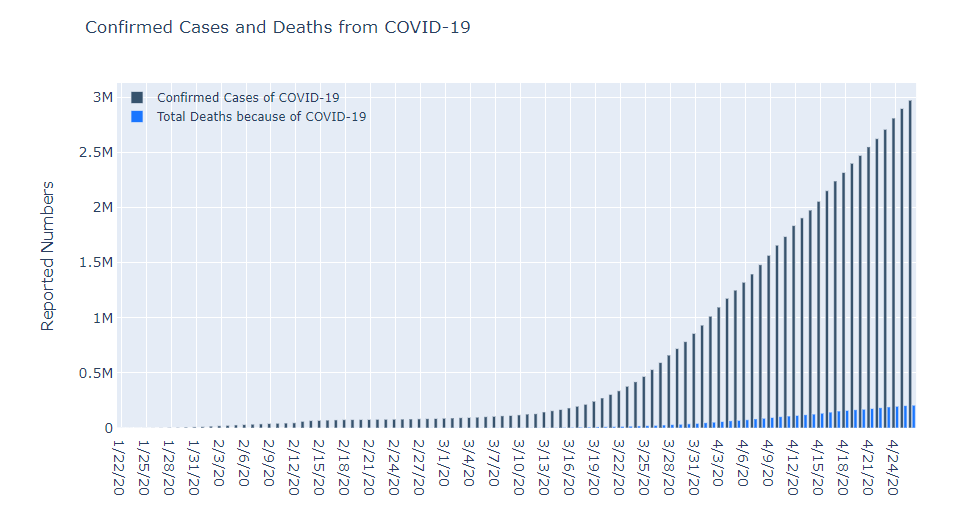


*Figure 14: Spread of deaths across the world*

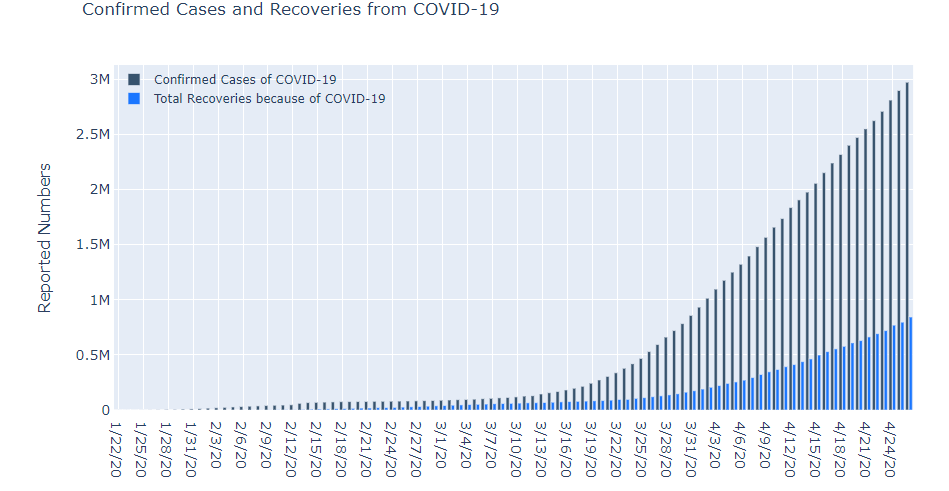
### 4.1.3 Bar charts

We also created two plots, one comparing the confirmed cases with deaths and one with confirmed cases and recoveries. In the first bar graph (Figure 15a) we notice that until 26/03 the number of confirmed cases is below half a million. For the next one month the number of cases constantly rises reaching 2 million on 15/04 and 3 million on 25/04. The number of deaths slightly increases compared to the infected cases, but doesn’t exceed 210.000.

Regarding the total numbers of recoveries from Covid-19 (Figure 15b) we can conclude that there are higher than the number of deaths in Figure 15a. Almost 1/3 of confirmed cases recovers. By 25/04 842.917 people have recovered from the virus.



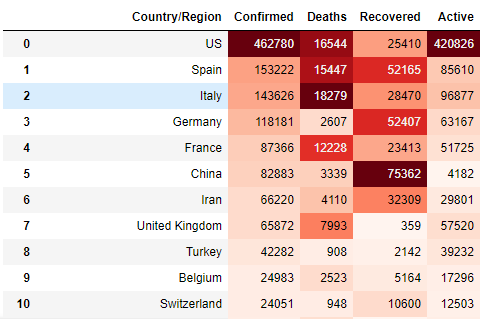
*Figure 15a: Number of confirmed cases and deaths according to date*



*Figure 15b: Number of confirmed cases and recoveries according to date*

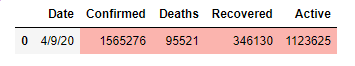
### 4.1.4 Tables and Choropleth map

We processed the data more and created an extra column with the active cases (Figure 16a). The column with the active cases is the result of confirmed cases excluding the deaths and recoveries. In Figure 16a we can see the top ten affected countries from Covid-19. Each column is depicted with shades of red according to the highest and lowest numbers in each case. We can conclude therefore that the US have the most confirmed and active cases with 462.780 cases and 420.826 active cases accordingly, Italy the highest number of deaths with 18.279, while China has the most recoveries with 75.362.



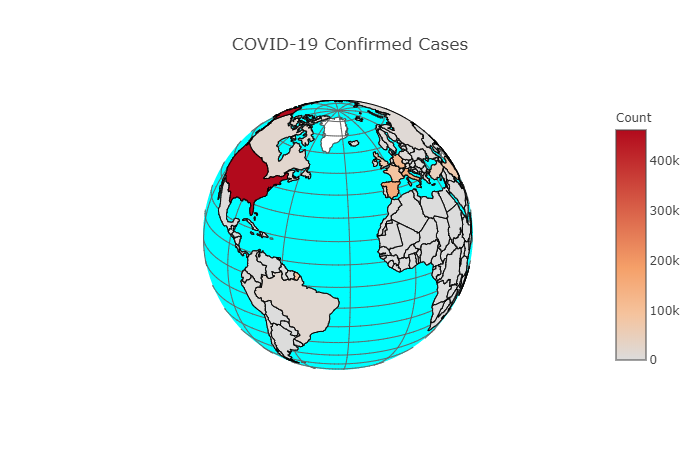
*Figure 16a: Table with top ten affected countries*

In Figure 16b we have the sum of all the confirmed cases, deaths, recoveries and active cases as of 09/04/2020. The confirmed cases reach 1.565.276 and the deaths 95.521.



*Figure 16b: Table with the sum of all cases*

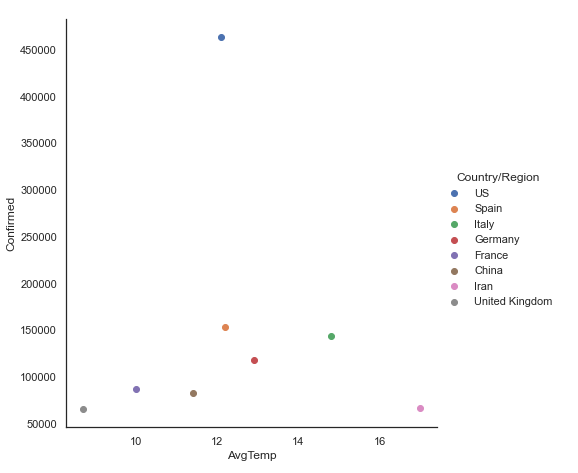
Moreover, we created a 3D map with the sum of confirmed cases in each country using pyplot and a choropleth map (Figure 17).



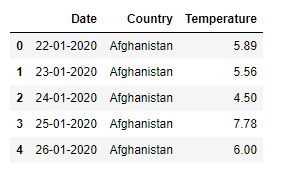
*Figure 17: Interactive map of total confirmed cases around the world*

### 4.1.5 Graphs related with temperature (change and add)

For the top 8 countries according to confirmed cases we found the average temperature for each one. Then we created a plot with confirmed cases and average temperature to see if there is any relationship between these two variables. In Figure 18 the US has the most cases with 450.000 people infected and the temperature is above 12oC. The rest of the countries don’t exceed 150.000 cases and the temperatures are between 8oC to 17oC.

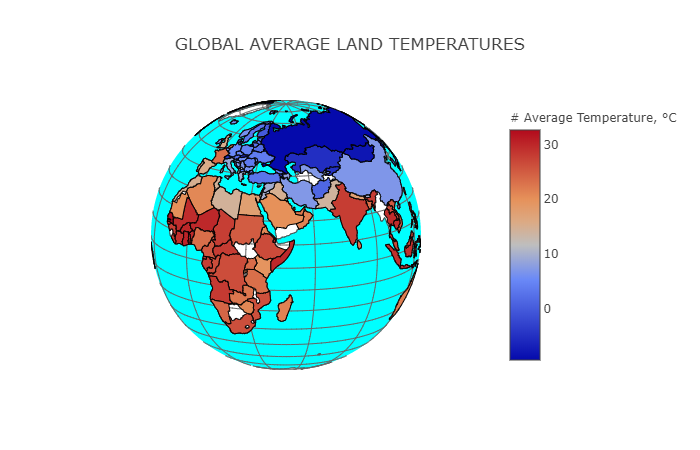
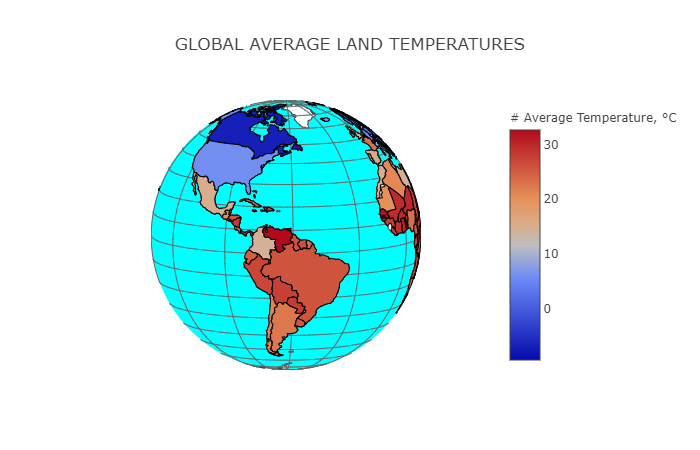


*Figure 18: Graph representing the top ten affected countries regarding temperature*



*Figure 19: Dataset that contains temperatures according to countries and dates*

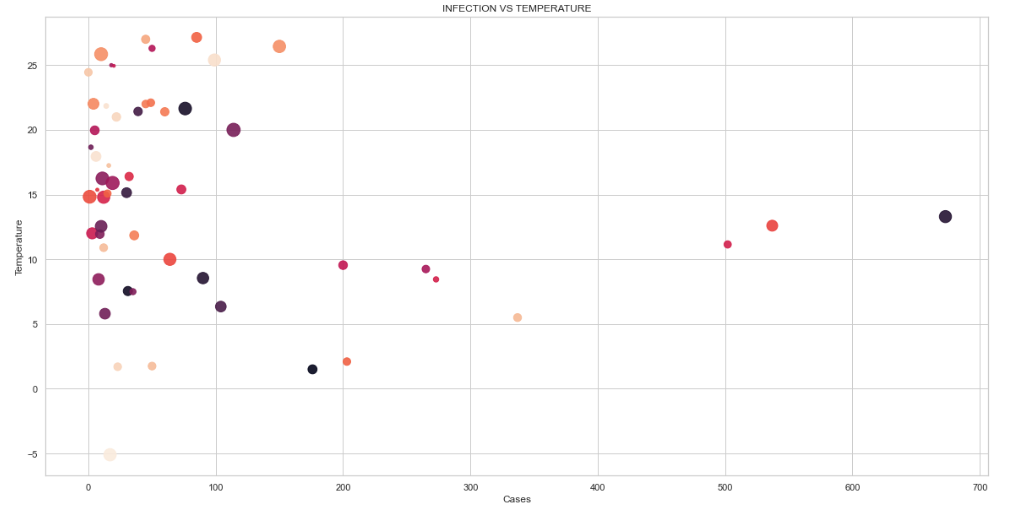
Using another dataset that contains the temperatures of all the countries we created a 3D map regarding the average temperature of each country (Figure19). From Figure 20 we can assume that during the outbreak and spread of the virus the countries most affected where the ones in the north where the average temperatures are pretty low regarding the south. These low temperatures are below 15o C.



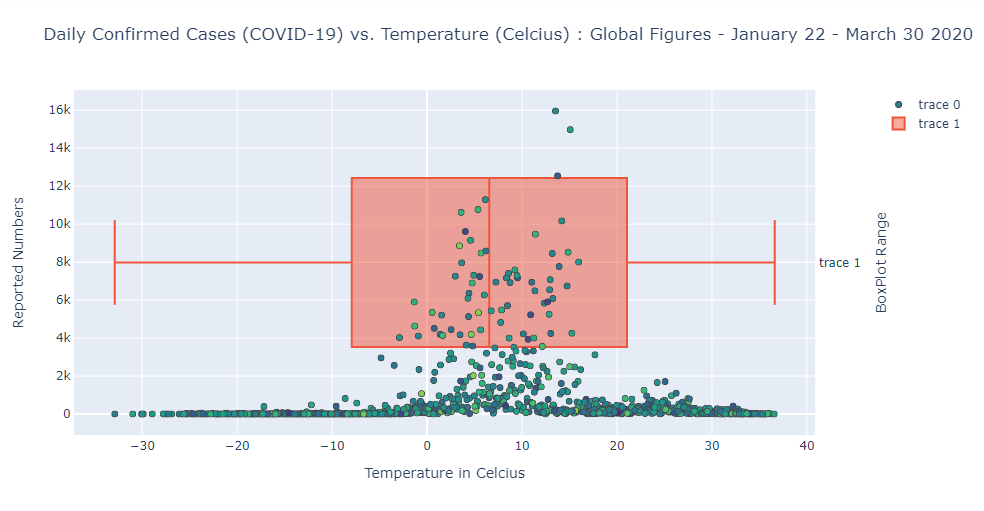
*Figure 20: Global temperatures according to country*

### 4.1.6 Relationship between temperature and confirmed cases

At the end we created a plot comparing the cases with the temperature to find a connection between them. In Figure 21 most confirmed cases can be identified between 5-17oC. (add more)



*Figure 21: Correlation between temperature and confirmed cases*



*Figure 22:*

## 4.2 Implementation of Data mining techniques

In this part we intent to implement various data mining techniques in order to identify any correlation between our variables, patterns and trends, as well as to predict future outcomes. Therefore, we are implementing Correlation and Time series forecasting techniques.

### 4.2.1 Correlation

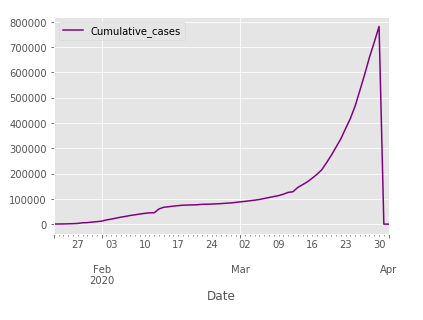
Correlation is a statistic that measures the degree of linear relationship between two quantitative variables. In this project we aim to examine the correlation between the number of cases around the world and various meteorological factors like temperature and humidity. For this purpose, we are going to use various libraries such as scipy, statsmodels and seaborn.

### 4.2.2 Time series forecasting

In this analysis we have the variable of time as a date. We can analyze this time series data in order to extract meaningful statistics and other characteristics. A time series is a set of observation taken at specified times usually at equal intervals. It is used to predict future values based on the previous observed values. By understanding past behavior we can forecast the future using an algorithm called time series analysis and also evaluate current accomplishments.

Before starting implementing any time series model we process our data. The variables that interest us are the dates and the number of cases according to date. Therefore, we will calculate the sum of all cases of all countries according to date and set the date as index.

Our dataframe consists of two columns: a date time index corresponding to each day since cases started been reported, and a count that corresponds to the number of cases during these months. In Figure 23 we have the plot of our time series.



*Figure 23: Number of cases according to date*

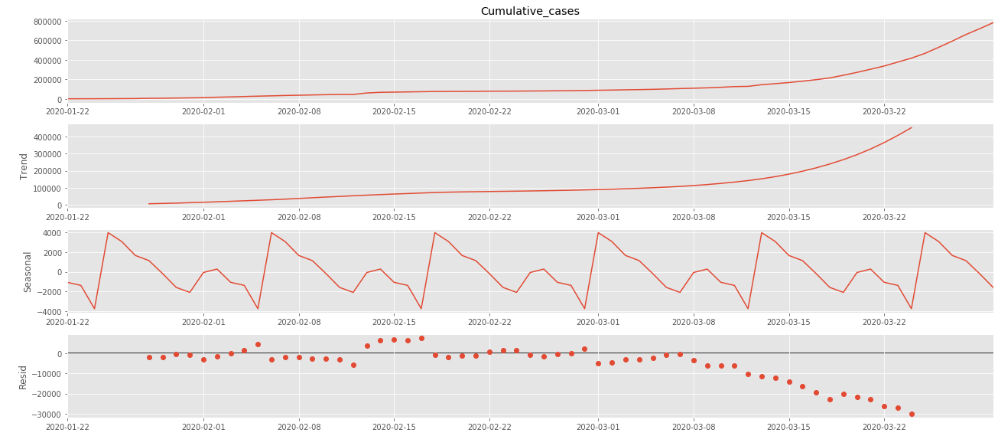
Plotting our time series reveals something interesting that would have been hard to notice earlier. We notice that the number of cases at the end of April look suspiciously low. This certainly appears to be a data problem. We remove all data after the 30th of April, since bad data will impact the accuracy of our model.

The trend shows the general tendency of our data to increase during a long period of time and the overall trend is upward.

To better understand the seasonal nature of our data, we can decompose our time series into components. The first step in decomposing our time series is determining whether our underlying stochastic process should be modeled with an additive or multiplicative decomposition. One heuristic here is if the magnitude of the seasonal fluctuations changes significantly over time, then use a multiplicative model. Otherwise, use an additive model. In our case, the magnitude of the seasonal fluctuations appears to be relatively consistent over time.

We will use a function in the statsmodels module to perform this decomposition for us, but we could compute it ourselves using a technique known as differencing. The four categories of the components of time series are:

* Trend
* Seasonal Variations
* Cyclic Variations
* Random or Irregular movements



*Figure 24:* *The cumulative cases (top) and its three additive components*

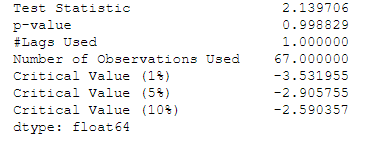
Figure 24 shows an additive decomposition of these data. The three components are shown separately in the bottom three panels of Figure 24. These components can be added together to reconstruct the data shown in the top panel. Notice that the seasonal component changes slowly over time. The remainder component shown in the bottom panel is what is left over when the seasonal and trend-cycle components have been subtracted from the data.

The grey bars to the right of each panel show the relative scales of the components. Each grey bar represents the same length but because the plots are on different scales, the bars vary in size. The large grey bar in the bottom panel shows that the variation in the remainder component is small compared to the variation in the data. If we shrunk the bottom three panels until their bars became the same size as that in the data panel, then all the panels would be on the same scale.

The seasonal\_decompose method generates this handy plot for us. And this plot helps highlight a few interesting things about our data. Firstly, it appears as though there has been an exponential growth in the spread of cases after January. Secondly, we notice that the peak of cases is at the end of April.

In the following step, we need to detect whether time series is stationary, and if not, we need to understand what kind of transformation is required to make it stationary. A time series is stationary when its statistical properties such as mean, variance, and autocorrelation are constant over time. In other words, time series is stationary when it is not dependent on time and not have a trend or seasonal effects. We can understand from the decomposition plot that due to trend and seasonality in our dataset we have a non-stationary time series.

We can apply statistical tests and Augmented Dickey-Fuller test is the widely used one. The null hypothesis of the test is time series has a unit root, meaning that it is non-stationary. We interpret the test result using the p-value of the test. If the p-value is lower than the threshold value (5% or 1%), we reject the null hypothesis and time series is stationary. If the p-value is higher than the threshold, we fail to reject the null hypothesis and time series is non-stationary.



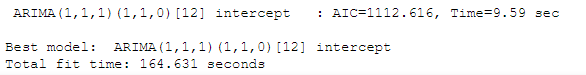
*Figure 25: Dickey-Fuller Test*

P-value is greater than the threshold value, we fail to reject the null hypothesis and time series is non-stationary, it has time dependent component (Figure 25). There are two major reasons behind non-stationary time series, trend and seasonality. We will apply differencing to make time series stationary by subtracting the previous observations from the current observations. Doing so we will eliminate trend and seasonality, and stabilize the mean of time series. Due to both trend and seasonal components, we apply one non-seasonal diff() and one seasonal differencing diff(12).

We are going to apply one of the most commonly used method for time-series forecasting, known as ARIMA, which stands for Autoregressive Integrated Moving Average. ARIMA models are denoted with the notation ARIMA(p, d, q).

We will use Python’s pmdarima library, to automatically extract the best parameters for our Seasonal ARIMA model. Inside auto\_arima function, we will specify d=1and D=1 as we differentiate once for the trend and once for seasonality, m=12 because we have monthly data, and trend='C'to include constant and seasonal=Trueto fit a seasonal-ARIMA. Besides, we specify trace=Trueto print status on the fits. This helps us to determine the best parameters by comparing the AIC scores.

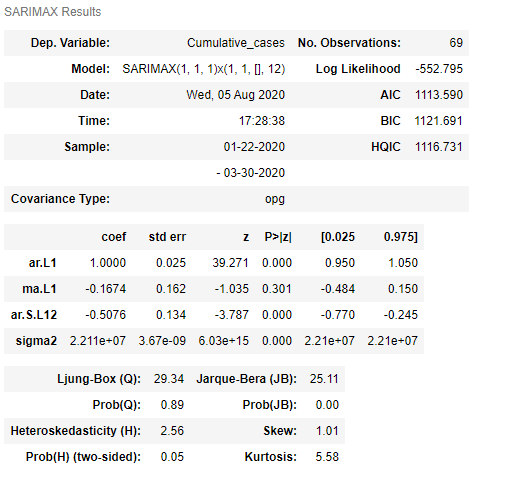
AIC (Akaike Information Criterion) is an estimator of out of sample prediction error and the relative quality of our model. The desired result is to find the lowest possible AIC score.



*Figure 26: AIC scores*

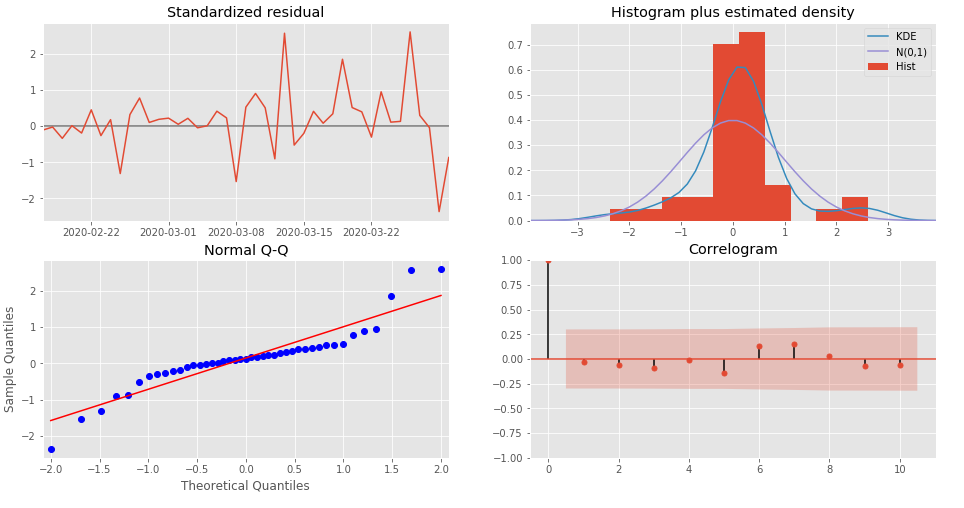
The result of auto\_arima function with various (p, d, q)(P, D, Q)m parameters indicates that the lowest AIC score is obtained when the parameters equal to (1, 1, 1)(1, 1, 0, 12) (Figure 26).

We create a SARIMA model by using SARIMAX function from statsmodel library. After fitting the model, we can also print the summary statistics (Figure 27).



*Figure 27: SARIMAX statistical results*

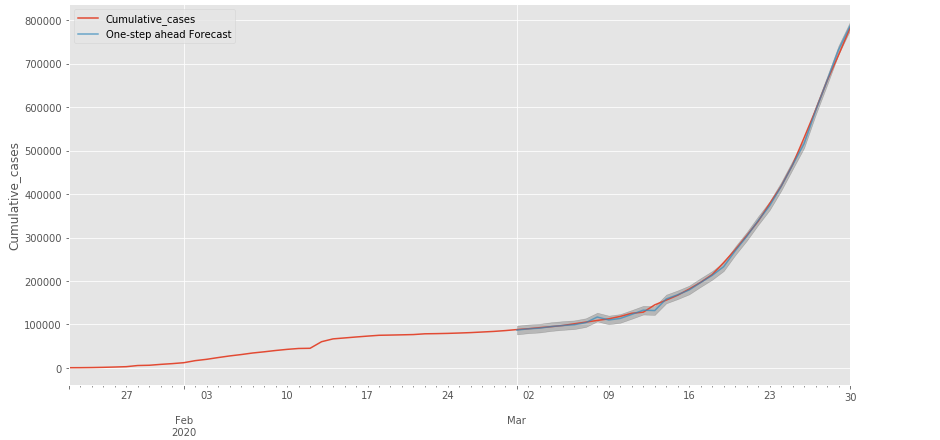
Primary concern of the model is to ensure that the residuals are normally distributed with zero mean and uncorrelated. To check for residuals statistics, we can print model diagnostics (Figure 28):

*Figure 28: Model diagnostics*

The top-left plot shows the residuals over time and it appears to be a white noise with no seasonal component. The top-right plot shows that kde line (in dark blue) closely follows the N(0,1) line, which is the standard notation of normal distribution with zero mean and standard deviation of 1, suggesting the residuals are normally distributed.

The bottom-left normal gg-plot shows ordered distribution of residuals (in blue) closely follow the linear trend of the samples taken from a standard normal distribution, suggesting residuals are almost normally distributed. The bottom-right is a correlogram plot indicating residuals have a low correlation with lagged versions. All these results suggest residuals are normally distributed with low correlation.

To help us understand the accuracy of our forecasts, we compare predicted cases to real cases of the time series, and we set forecasts to start at 2020–03–01 to the end of the data.

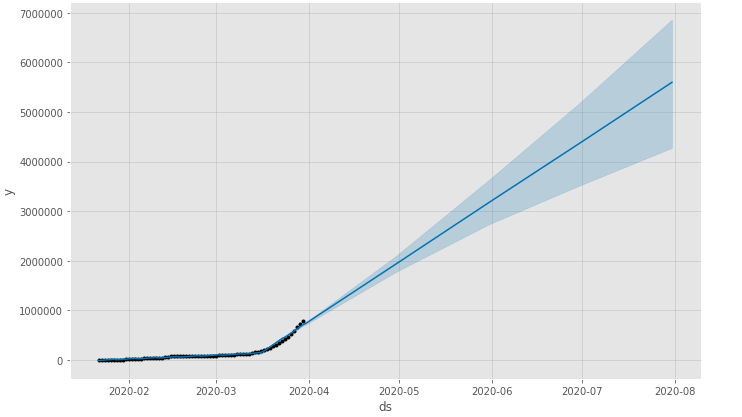


*Figure 29:* *Plot of the observed values compared to the rolling forecast predictions.*

The line plot is showing the observed values compared to the rolling forecast predictions. Overall, our forecasts align with the true values very well, showing an upward trend starts from the beginning of March.

The above procedure gives us a great overall picture of the data, but we’d like use the historical data to forecast future cases and we’ll use Prophet to do that.

Prophet is a module that enables time-series forecasting. Prophet uses an additive decomposable time series model. In a Prophet model Yt=g(t)+s(t)+h(t)+ϵt, there are three main components: a trend function g(t), a seasonality function s(t), a holidays function h(t), ϵt is an error function.



*Figure 25: Forecast model of cases until August*

Prophet forecasts that we will see a continuation of the upward trend in the spread of cases that has begun in March (Figure 25). The plotted forecast includes our actual data points, as well as the forecast on the future. This allows us to see where our actual observed data lie outside of our uncertainty level.

Compare it with dataset I find until august

# Chapter 5: Conclusion – Future Research

# Bibliography

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., … others. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet*, *395*(10223), 507–513.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Larose, D. T., & Larose, C. D. (2014). *Discovering knowledge in data: an introduction to data mining* (Vol. 4). John Wiley & Sons.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., … others. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus--infected pneumonia. *New England Journal of Medicine*.

Lu, H., Stratton, C. W., & Tang, Y.-W. (2020). Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle. *Journal of Medical Virology*, *92*(4), 401–402.

Rothe, C., Schunk, M., Sothmann, P., Bretzel, G., Froeschl, G., Wallrauch, C., … others. (2020). Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. *New England Journal of Medicine*, *382*(10), 970–971.

Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., … others. (2020). Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus--infected pneumonia in Wuhan, China. *Jama*, *323*(11), 1061–1069.