Contents

[1. Introduction 3](#_Toc149816117)

[2. Motivation 3](#_Toc149816118)

[3. Business Understanding 3](#_Toc149816119)

[5. Business Description 3](#_Toc149816120)

[5.1 Research Question 3](#_Toc149816121)

[5.2 General goal 3](#_Toc149816122)

[5. Success criteria/indicators 4](#_Toc149816123)

[6. Technologies used 4](#_Toc149816124)

[6.1 Models and machine learning algorithms 4](#_Toc149816125)

[6.2 Libraries 4](#_Toc149816126)

[6.3 Methodology for used for the Project Management Framework 4](#_Toc149816127)

[6.4 Accomplishment Data 4](#_Toc149816128)

[7. Data 4](#_Toc149816129)

[7.1 Characterization of the dataset 4](#_Toc149816130)

[7.1.1 Attributes 4](#_Toc149816131)

[7.1.2 Dimensions 5](#_Toc149816132)

[7.1.3 Descriptive Statistics 5](#_Toc149816133)

[7.2 Data Preparation and Pre-processing 5](#_Toc149816134)

[7.2.1 Dropping Duplicates 5](#_Toc149816135)

[7.2.2 Missing Values 5](#_Toc149816136)

[7.3 Data Visualization 6](#_Toc149816137)

[7.3.1 Confirmed Cases Chart 6](#_Toc149816138)

[7.3.2 Death Cases Charts 9](#_Toc149816139)

[7.3.3 Recovered Cases Charts 11](#_Toc149816140)

[7.3.4 Active Cases Charts 13](#_Toc149816141)

[8. Normalizing the data 15](#_Toc149816142)

[9. Machine Learning Models 15](#_Toc149816143)

[10. Dimensionality Reduction 18](#_Toc149816144)

[10.1 PCA 18](#_Toc149816145)

[11. Feature Engineering 19](#_Toc149816146)

[11.1 Applying the models with 10% testing and 90% training 19](#_Toc149816147)

[Hyperparameter tuning with KNeighbors Regressor 19](#_Toc149816148)

[Hyperparameter tuning with Random Forest Regressor 19](#_Toc149816149)

[11.2 Applying the models with 20% testing and 80% training 19](#_Toc149816150)

[Hyperparameter tuning with KNeighbors Regressor 19](#_Toc149816151)

[Hyperparameter tuning with Random Forest Regressor 20](#_Toc149816152)

[11.3 Applying the models with 30% testing and 70% training 20](#_Toc149816153)

[Hyperparameter tuning with KNeighbors Regressor 20](#_Toc149816154)

[Hyperparameter tuning with Random Forest Regressor 20](#_Toc149816155)

[12. Results 20](#_Toc149816156)

[13. Conclusion 20](#_Toc149816157)

[14. Appendix 20](#_Toc149816158)

[14.1 Appendix 1: Data Dictionary 20](#_Toc149816159)

[14.2 Appendix 2: CRISP-DM - Part 1 21](#_Toc149816160)

[14.3 Appendix 3: CRISP-DM - Part 2 21](#_Toc149816161)

[15. Team Collaborations 21](#_Toc149816162)

[15.1 Mijail Blanco’s Collaboration 21](#_Toc149816163)

[15.2 Emily Herbas Collaborations 21](#_Toc149816164)

[16. References 24](#_Toc149816165)

# 1. Introduction

The COVID-19 pandemic has had an impact on the world and has led to significant changes in our daily lives, economies, and healthcare systems.

A great example is the economic slowdown that has affected businesses and jobs. Many companies increased the use of digital technology to adapt to the crisis. However, many companies were also forced to reduce salaries, work hours, and lay off workers.

In the healthcare sector, the pandemic led to an increased need for effective, accessible, and affordable medical care, which in many countries was not possible due to the high number of daily infected cases. This is why many countries chose to implement quarantines to reduce the number of infected cases.

The dataset we are going to analyse contains information on confirmed cases, deaths, recoveries, and active cases worldwide. This information will help us visualize and understand what was previously mentioned.

# 2. Motivation

The principal motivation for this analysis is to apply what we have been learning in Machine learning into the analysis of Covid cases, making the predictions of the confirmed cases of Covid during the very beginnings of the pandemic in Europe Region. To know the number of positive cases was primordial since there were many people affected that needed medical services and treatments.

# 3. Business Understanding

To predict confirmed Covid cases, we are going to apply different Machine Learning models, which data was chosen from a Covid repository where we can analyse the beginning of Covid cases in the world, but we decided just to focus in Europe region for making better predictions.

# 5. Business Description

## 5.1 Research Question

Using different training and tests splits in the data, what model could predict better the confirmed cases of Covid in Europe (WHO Region)?

# 5.2 General goal

The general goal is to predict how many confirmed Covid cases. This will help European countries to know how the virus is spreading for taking preventive, and medical decisions in the European Region.

# 5. Success criteria/indicators

We are going to apply 3 different Machine Learning Models in which we are going to compare the R-squared value of each or the Mean Absolut Error to define which is the best model.

# 6. Technologies used

## 6.1 Models and machine learning algorithms

As we are trying to predict a numerical value, we are going to apply supervised Models for Regression that fit with our data and we decided to use (KNeighbors Regressor and Random Forest Regressor) for regression, and ARIMA model for time series analysis.

## 6.2 Libraries

We used different libraries to perform this analysis like Pandas, Seaborn, Matplotlib, NumPy, Standard Scaler, PCA, ARIMA, sm, adfullet, among others.

## 6.3 Methodology for used for the Project Management Framework

We decided to used CRIP-DM methodology in which we defined the data to analyse, we developed the Business and Data understanding, Data Preparation, Modelling, and the Presentation of the Results which you can see in Appendix 2.

As it is stated by IBM (www.ibm.com, n.d.) CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining, is an industry-proven way to guide your data mining efforts.

-As a methodology, it includes descriptions of the typical phases of a project, the tasks involved with each phase, and an explanation of the relationships between these tasks.

-As a process model, CRISP-DM provides an overview of the data mining life cycle.

## 6.4 Accomplishment Data

Our Covid Dataset was split through different excel files and we decided to concatenate with the useful information for our project and after that we have 49068 rows and 11 columns in the dataset in which 2 of them are continuous numerical variables, 4 are discrete numerical variables and 5 are categorical variables. We have data collected from February 2020 until July 2020, confirmed, death, recovered cases of Covid and other variables that are available in the Data Dictionary (Appendix 1)

**Source**

The data was chosen from a Kaggle repository found in this link: <https://www.kaggle.com/datasets/imdevskp/corona-virus-report?select=covid_19_clean_complete.csv> (Kaggle, 2020)

# 7. Data

## 7.1 Characterization of the dataset

### 7.1.1 Attributes

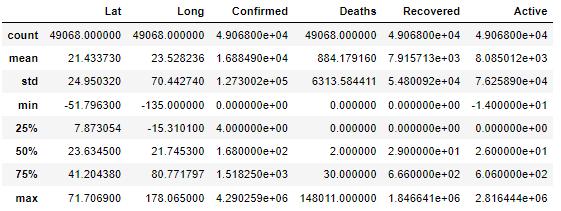
We are going to analyse 11 variables in which “Confirmed” is going to be our target variable to be predicted, and the others are going to be independent variables which will be selected for our analysis.

### 7.1.2 Dimensions

The shape of the Covid dataset to be analysed is 49068 rows and 11 columns.

### 7.1.3 Descriptive Statistics

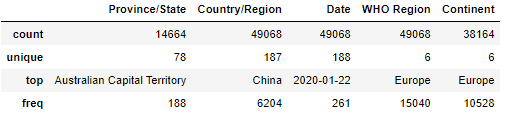
In Figure 1 we are going to see the principal statistics of the numerical variables.

****

*Figure 1 Statistics of the numerical variables in Covid dataset*

In the dataset we can appreciate some details as latitude, and longitude which show us where is every country located, and a collection of confirmed, deaths, recovered, and active cases of Covid.

We are going to analyse the statistics of the categorical variables in Figure 2



*Figure 2 Statistics of the categorical variables in Covid dataset*

In our categorical variables, we can see that we have 187 countries in our dataset, 78 provinces, 188 different dates, 6 WHO Region and 6 different continents

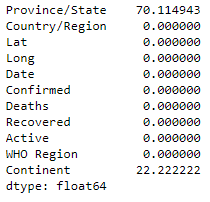
## 7.2 Data Preparation and Pre-processing

### 7.2.1 Dropping Duplicates

After dropping duplicates, the number of rows didn't change after dropping duplicates, and this means that our data doesn't have duplicates inside.

### 7.2.2 Missing Values

The NaN values were standardized and we got the next results analyzing missing values:



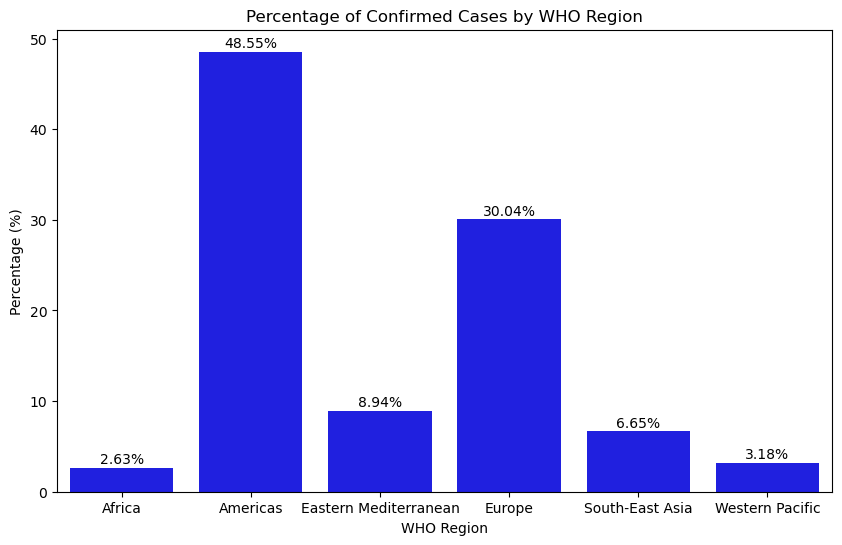
*Figure 3: Percentage of NaN values in each column*

As we can see the variable "Province/state" has around 70% of Null values present, and we decided to drop it since it would create bias and variance leading into bad predictions in our Machine Learning predictions, and in the column “Continent”, we explored that "WHO Region" and “Continent”, we have similar unique values; however, they are not the same for example in Eastern Mediterranean Regions could be included European and non/European countries, that is why we are going to replace the NAN- values as "Unknown".

## 7.3 Data Visualization

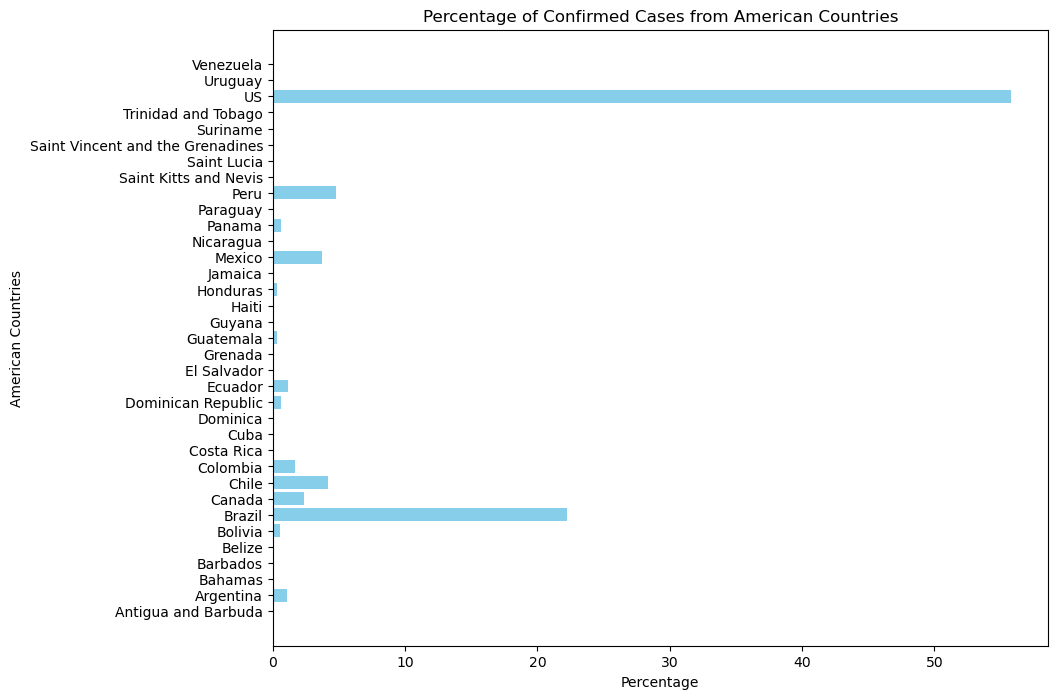
In this part we want to visualize by region defined by the World Healt Organization (WHO Region) the percentage of people in the different cases of Covid (Confirmed, Active, Death and Recovered) in order to see which Region was the most affected.

### 7.3.1 Confirmed Cases Chart



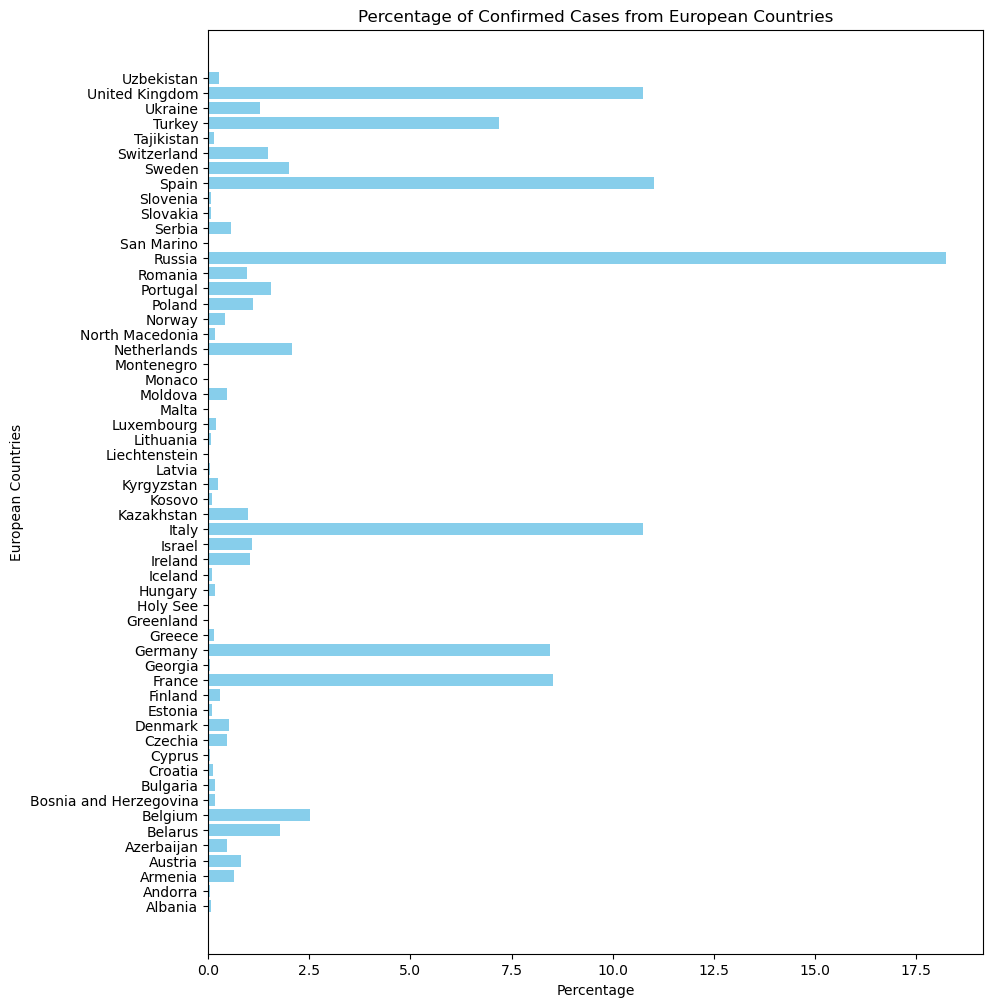
*Figure 4: Percentage of Confirmed Cases by WHO Region*

We can observe that the Americas WHO Region has the highest number of confirmed cases with 48,55% followed by Europe with 30,04%.



*Figure 5: Percentage of Confirmed Cases from American Countries*

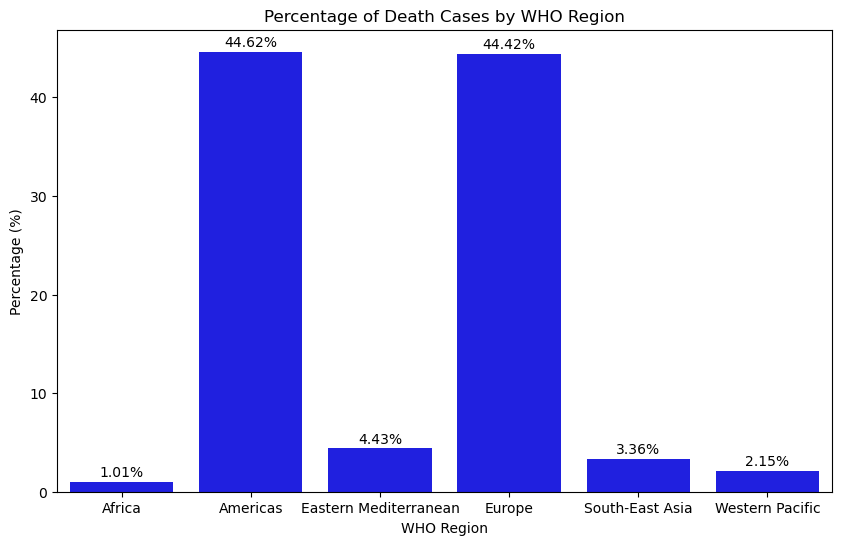
We can observe that US with more than 55% and Brazil with around 22% has the highest number of confirmed cases.



*Figure 6: Percentage of Confirmed Cases from European Countries*

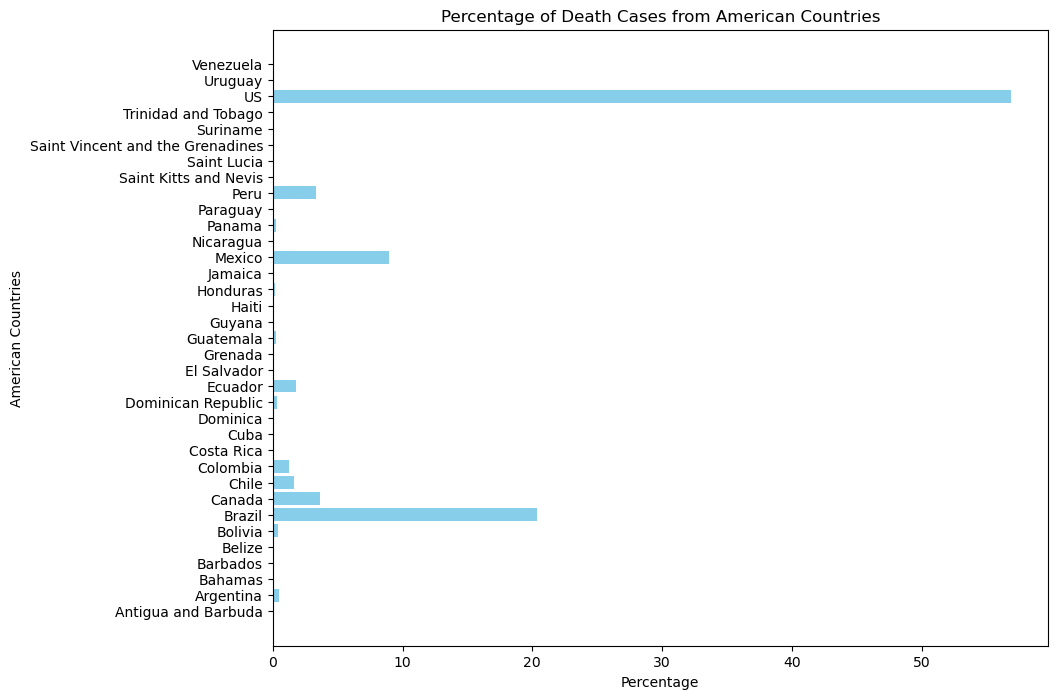
We can observe that Russia with 17,5% has the highest percentage of confirmed cases followed by Spain, United Kingdom and Italy (with 11,5%).

### 7.3.2 Death Cases Charts



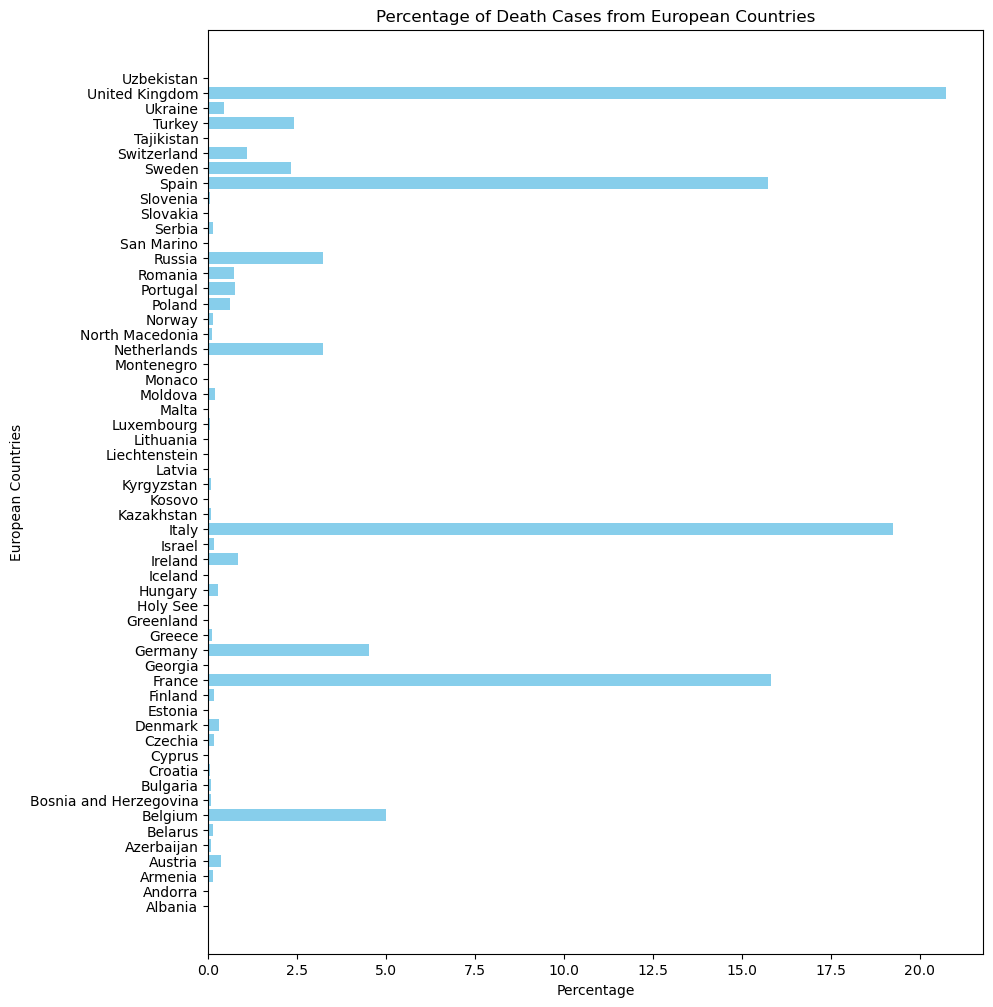
*Figure 7: Percentage of Death Cases by WHO Region*

In the previous graph, we can observe that the Americas with 44,62% and Europe with 44,42% have the highest number of death cases. To identify which country or countries in both WHO Regions have the highest number of death cases we will generate new bar charts.



*Figure 8: Percentage of Death Cases from American Countries*

We can see that the US with more than 55% has the highest percentage of deaths followed by Brazil with 20%, this may be due to the fact that both countries have a larger population compared to the rest of the countries in the Americas.

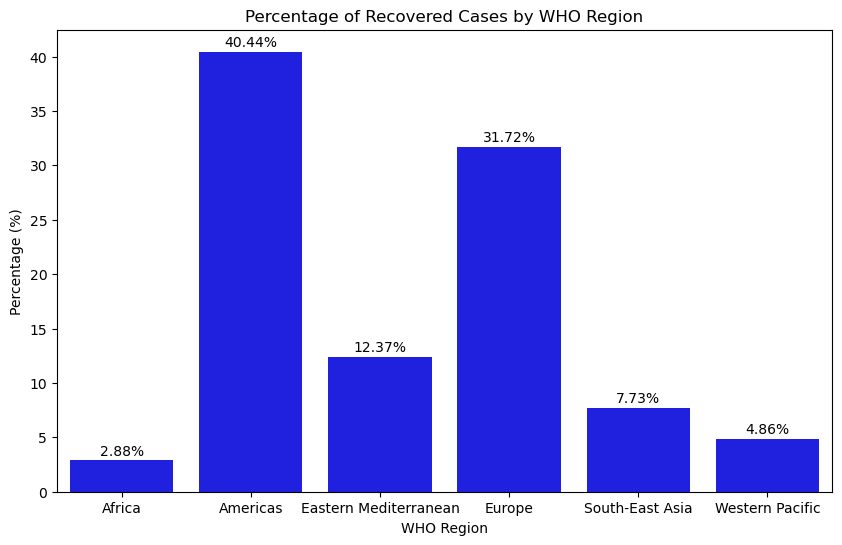


*Figure 9: Percentage of Death Cases from European Countries*

In this case, we can observe that UK with a 20% has the highest number of COVID-related deaths, followed by Italy (around 18,5%), Spain (16.5%) and France (16,5%). This leads us to understand:

We can suppose that maybe these countries were not equipped to handle infected patients, as reflected in the fact that it enforced one of the longest home quarantines for its population but also, we can see that the percentages are not as big as US and Brazil.

### 7.3.3 Recovered Cases Charts



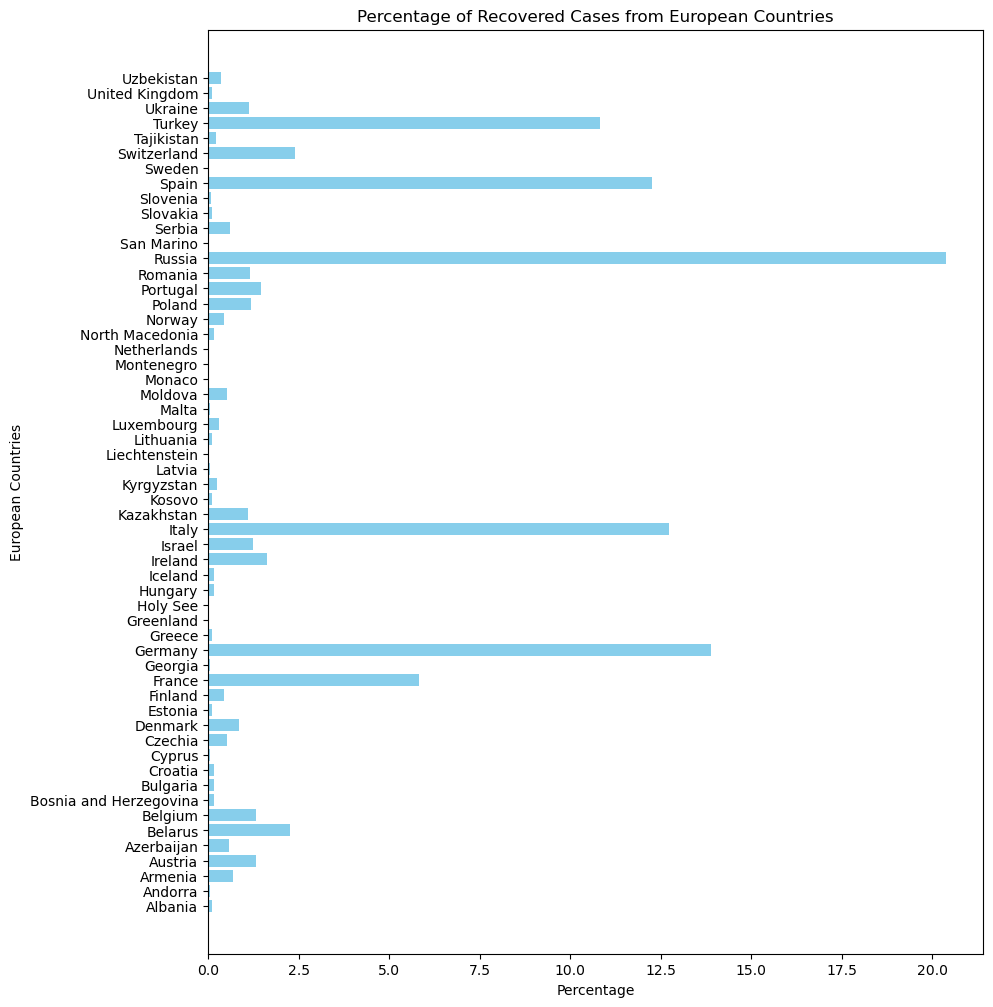
*Figure 10: Percentage of Recovered Cases by WHO Region*

We can observe that the Americas WHO Region has the highest number of Recovered cases with 40,44% followed by Europe with a 31,72%. To identify which country or countries in the Americas has the highest number of recovered cases, we will generate a new bar chart



*Figure 11: Percentage of Recovered Cases from American Countries*

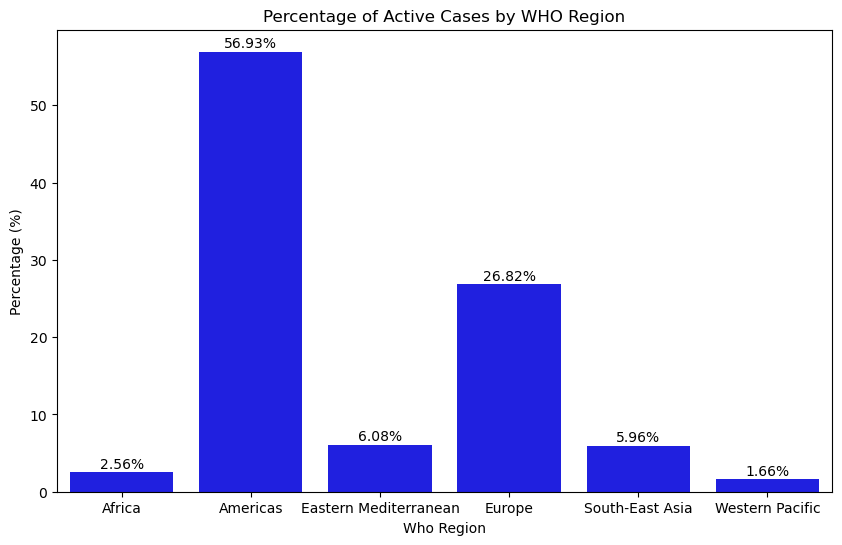
In this case we can observe that US with 35% and Brazil with 34,8% have the highest percentage of cases Recovered.



*Figure 12: Percentage of Recovered Cases from European Countries*

We can see that Russia has the highest percentage of recovered cases with a 20% followed by Germany (around 14%), Italy (12,5%), Spain (11,5%) and Turkey (11,5%).

### 7.3.4 Active Cases Charts



*Figure 13: Percentage of Active Cases by WHO Region*

The Americas Who Region has the highest percentage of Active cases with 56,93%, although Europe had percentages very close to the Americas in the previous graphs, we can now see that the percentage of active cases is 26.82%, less than half of the cases in the Americas.

We could say that Europe at some point was able to manage Covid infections, a possible cause for the reduction in cases may be the long-term confinement that was imposed in some countries.

To identify which country or countries in the Americas have the highest number of active cases we will generate a new bar chart.

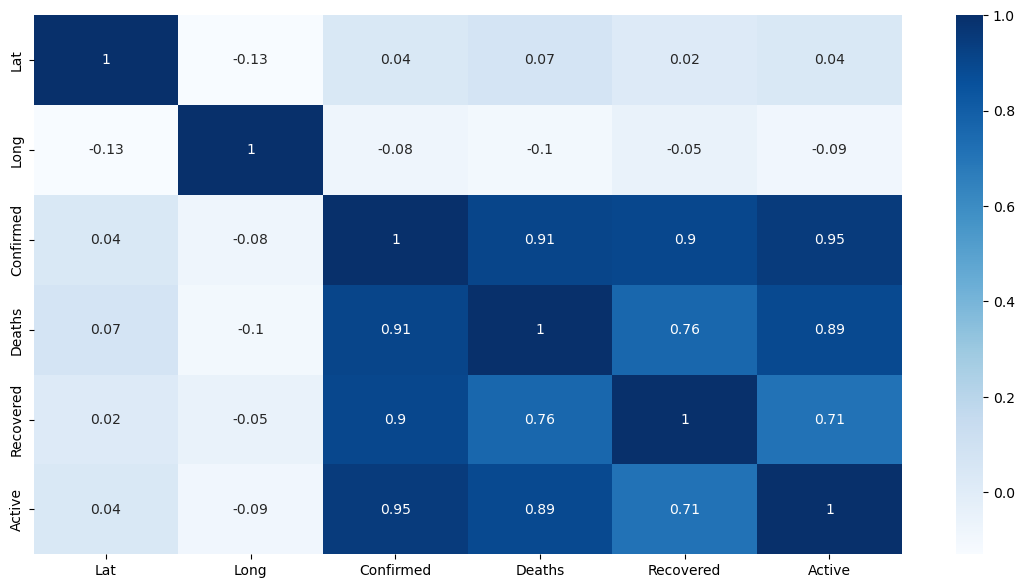


*Figure 14: Percentage of Active Cases from American Countries*

We can see that the United States has the highest percentage of active cases at nearly 70%. This could mean:

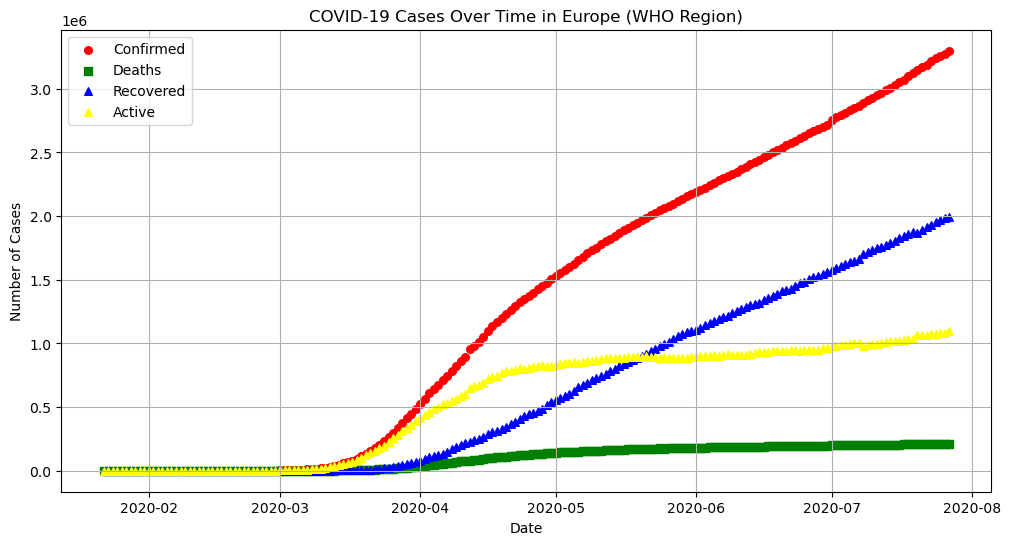
* The population did not fully adhere to the precautionary measures implemented by the country.
* Due to its status as a major tourist destination, there may not have been strict restrictions on entry from other countries to avoid negatively impacting the country's economy.
* Disruptions to schools and businesses may have been kept to a minimum.

**Correlation:**



*Figure 4: Correlation Matrix between the numerical values of Covid cases*

In Figure 4, in the Correlation Matrix of the numerical values, we can see a strong correlation of the target variable “Confirmed” with “Deaths”, “Recovered” and “Active”, these correlations are really good since they will help us get good Machine Learning models for predictions.



*Figure 5: Visualization of Covid 19 Cases Over Time in Europe (“WHO Region)*

In this Figure 5, we can see a positive tendency trend in “Confirmed” Covid cases, but also, we can see how was the distribution of the Covid cases Over time in Europe according to our data time provided. We can see also that until march the cases were still around 0. Besides, we can appreciate how was the tendency of Deaths, Recovered, and Active cases in Europe Region.

# 8. Normalizing the data

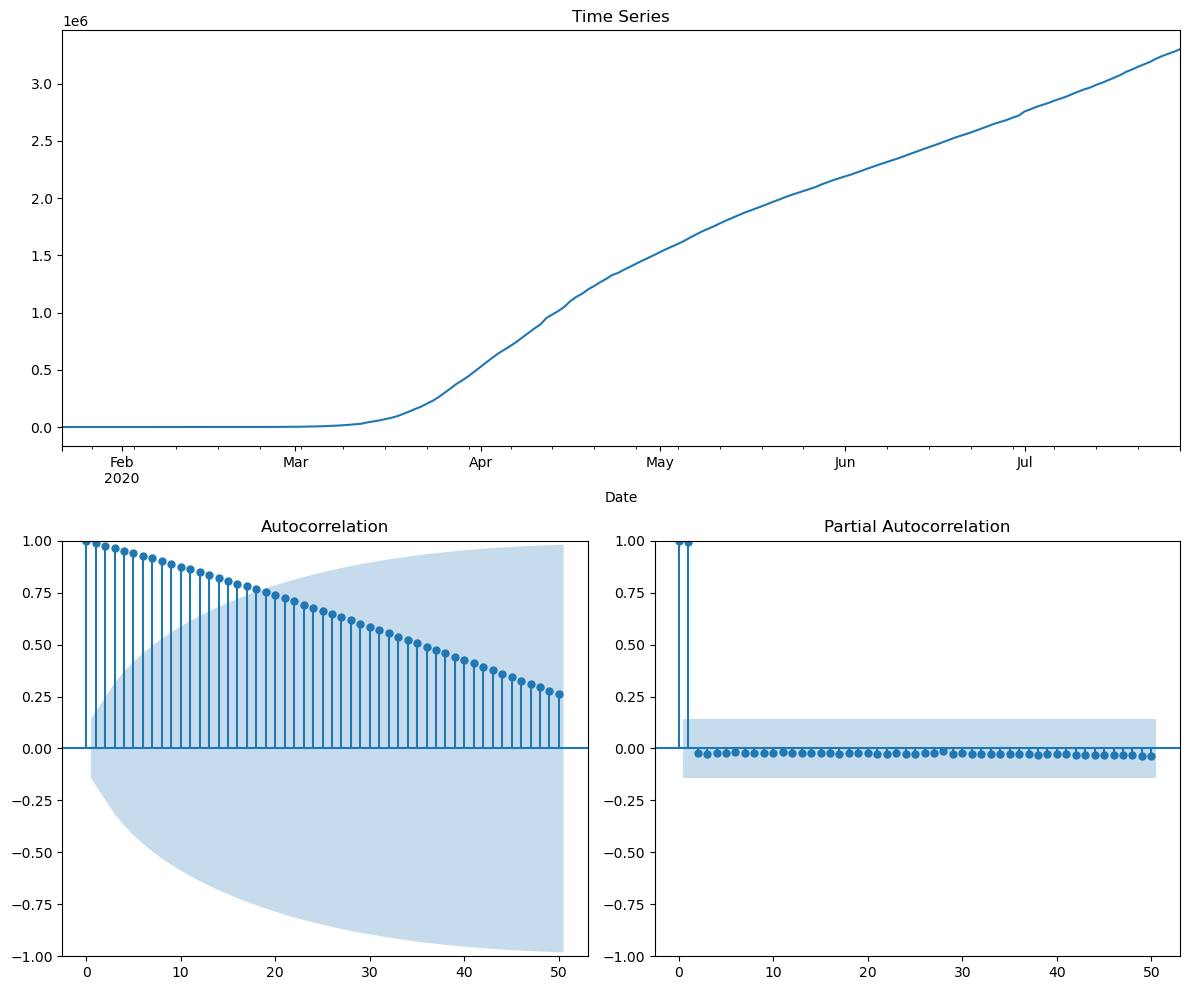
For the time series model we didn’t normalize the data since we have only one column to analyse. However, in KNeighbord Regressor and Random Forest Regressor, we used all the columns of the dataset and we tried to normalize with Standard Scaler, Minmax Scaler and Robust Scaler, but our machine learning results give around 100% of accuracy and that is why we decided not to scale the data for the following Machine Learning Models.

# 9. Machine Learning Models

**A) Arima Time Series Model:**

According to Eryc (2020, p.111) ARIMA models are a class of statistical models that are used for analyzing and forecasting time series data. They aim to do so by describing the autocorrelations in the data. ARIMA stands for Autoregressive Integrated Moving Average and is an extension of a simpler ARMA model. The goal of the additional integration component is to ensure stationarity of the series, because, in contrast to the exponential smoothing models, the ARIMA class requires the time series to be stationary. In the next few paragraphs, we briefly go over the building blocks of ARIMA models.

The first Machine Learning model analysed is ARIMA model in which we applied techniques learned in class to make predictions over time, as our data is very reduced of just 7 months, we decided not to apply SARIMAX since Covid cases didn’t have too much peaks during this time, following a continuous increasing tendency.



*Figure 6: Visualization of Covid 19 Cases Over Time in Europe (“WHO Region)*

We did the previous plot (Figure 6) to see if our data is shown as stationary, but we can see that our data is not stationary since it has an increasing trend in the first visualization where we see the months vs confirmed Covid cases, we can also see a decreasing autocorrelation and analyzing the partial autocorrelation, we can see around the first 2 lags are around 1 in the y-axes, but then we can see the following lags in negative values close to zero and we can see that there is a relationship present there.



*Figure 7: Results of the mean and variance values of the first 110 rows and afterwards*

We can see that the variance is changing between the first 110 rows and afterwards and we can say that the data is not stationary, and to confirm this using a mathematical method, we will use Dickey Fuller Test (ADF).

SPUR ECONOMICS (SPUR ECONOMICS, 2023) mentions that The Dickey Fuller Test is a unit root based test of stationarity. The unit root based tests focus on the coefficient associated with the first lag of the time series variable. If the coefficient is one (has a unit root), the time series behaves similarly to a Random Walk model which is non-stationary. Hence, we can statistically test whether that coefficient is equal to one. The Dickey Fuller Test adopts this procedure by carefully manipulating equations to test for stationarity.

For the Dickey Fuller test, we stated the next Hypothesis to compare that is the null hypothesis (H0), and the alternative one (HA):

H0: The time series is non-stationary

HA: The time series is non-stationary

The significance level stablished by default is 0.05



*Figure 8: ADF and p-value results applying Dickey Fuller Test*

Applying Dickey Fuller Test, we can see that the p value is more than 0.05; so, we reject the null hypothesis, and we can say that our original data is non-stationary.

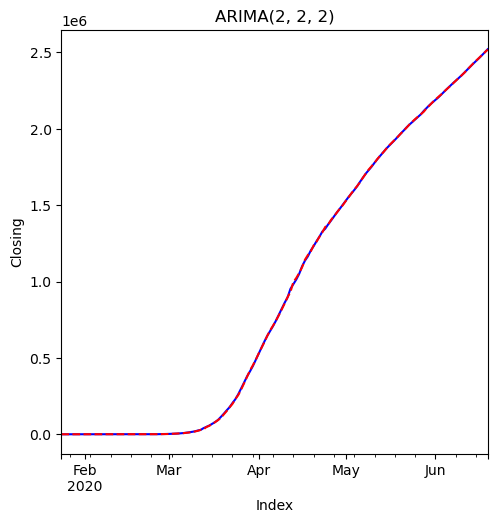
**Splitting the data and calculating metrics:**

Before applying techniques to transform the data stationary, we did 3 different splits for the analysis in ARIMA Time Series Model and we got the next values for a good performance of the model:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Test Split** | | |
|  | **20%** | **15%** | **10%** |
| **Number of differences to make it stationary** | 13 | 31 | 22 |
| **New p-value** | 0.047846 | 0.049262 | 0.003939 |
| **Best p, d, q values** | (2,2,2) | (2,2,2) | (0,2,1) |
| **AIC** | 2846.044 | 3004.792 | 3246.212 |
| **BIC** | 2861.030 | 3020.074 | 3252.448 |
| **Train R-Squared** | 0.999986 | 0.999988 | 0.999985 |
| **Test R-Squared** | 0.995756 | 0.803456 | 0.885761 |
| **Mean Absolut Error (MAE)** | 9713.731314 | 67070.543526 | 34248.144607 |

*Table 1: Results from ARIMA model with different test splits.*

In Table 1, we can appreciate the number of differences applied for making the data stationary calculating the new p-value with Dickey Fuller test with the same hypothesis stated previously. We analysed the recovered cases adding p, d, and q values iterating from initial values in which p is 1 which means the autoregression, 1 as q that is the difference and 0 as moving average. After that we applied a function to determine which is the minimum AIC value with different iterations. AIC tells us how accurate we are in our model with (p,d,q) combination, and BIC give us information criterium about the dataset, and in the next code we will train again our ARIMA model. We can see difference between R-Squared metric comparing the train and test with different splits, and we also calculated the Mean Absolut Error to compare between them showing that the lowest error is found in 20% test split compared to the other models.



*Figure 7: ARIMA model with 20% test and optimal p, d, q values*

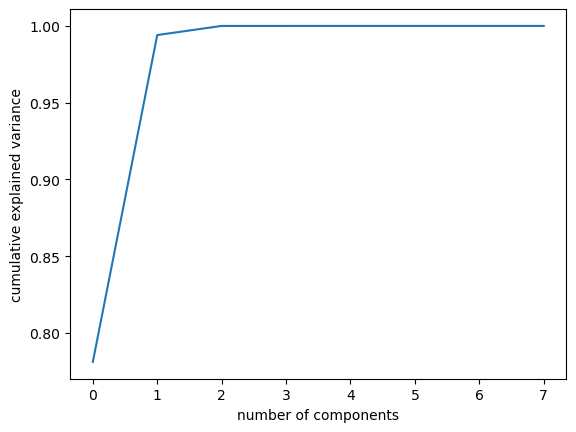
In Figure 7, we can observe how our ARIMA model fits well when we apply 20% of testing and we have p, d, q values as (2, 2,2).

# 10. Dimensionality Reduction

## 10.1 PCA

As we know the PCA is an unsupervised technique used to reduce the dimensionality and for identify trends in high-dimensionality sets.

In this case we applied the PCA for the Confirmed cases and we reduce our columns from 8 to 3.



# 11. Feature Engineering

We applied the Hyperparameter tuning with the K-Neighbords Regressor and Random Forest Regressor models.

The Hyperparameter tuning is a process in training machine learning models that can significantly enhance the model’s performance and this can help to have better prediction accuracy.

## 11.1 Applying the models with 10% testing and 90% training

## Hyperparameter tuning with KNeighbors Regressor



## Hyperparameter tuning with Random Forest Regressor



## 11.2 Applying the models with 20% testing and 80% training

## Hyperparameter tuning with KNeighbors Regressor



## Hyperparameter tuning with Random Forest Regressor



## 11.3 Applying the models with 30% testing and 70% training

## Hyperparameter tuning with KNeighbors Regressor



## Hyperparameter tuning with Random Forest Regressor



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Test Split** | | | | | |
|  | **10%** | | **20%** | | **30%** | |
| **Model** | **Kneighbors Regressor** | **Random Forest Regressor** | **Kneighbors Regressor** | **Random Forest Regressor** | **Kneighbors Regressor** | **Random Forest Regressor** |
| **Mean Absolute Error** | 40,62 | 114,49 | 0 | 0 | 0 | 0 |
| **Accuracy** | 0,99 | 0,99 | 1 | 1 | 1 | 1 |

We can observe that with 10% testing and 90% training, the values of the Mean Absolute Error (MAE) are high in both models. While we have very high accuracy, we must also consider that for regression tasks, accuracy is not the most common evaluation metric. In this case, the specific regression metric would be the value obtained in the MAE.

# 12. Results

# 13. Conclusion

# 14. Appendix

## 14.1 Appendix 1: Data Dictionary

|  |  |
| --- | --- |
| **Columns** | **Description** |
| Province/State | This is the name of the Province/State associated with the Covid-19 data in the dataset. |
| Country/Region | This is the name of the Country or geographic region associated with the Covid-19 data in the dataset. |
| Lat | Latitude of location |
| Long | Longitude of location. |
| Date | Date of cumulative report. |
| Confirmed | Cumulative number of confirmed cases. |
| Deaths | Cumulative number of death cases. |
| Recovered | Cumulative number of recovered cases. |
| Active | Cumulative number of active cases. |
| WHO Region | A region defined by the World Health Organization (WHO) that includes that country or region within the framework of that organization. |
| Continent | Continent on which the country or region is located. |

## 14.2 Appendix 2: CRISP-DM - Part 1

## 14.3 Appendix 3: CRISP-DM - Part 2

# 15. Team Collaborations

## 15.1 Mijail Blanco’s Collaboration

## 15.2 Emily Herbas Collaborations

The first contribution that I made in this project was to search for the data, this data had to be at least 8000 rows ant 10 columns after the cleaning, the data was founded in Keaggle.

Since our initial data came in different .csv documents, what was done as a team was to analyse each document and see which columns would be useful for the analysis and thus be able to concatenate the documents and clean the data correctly.

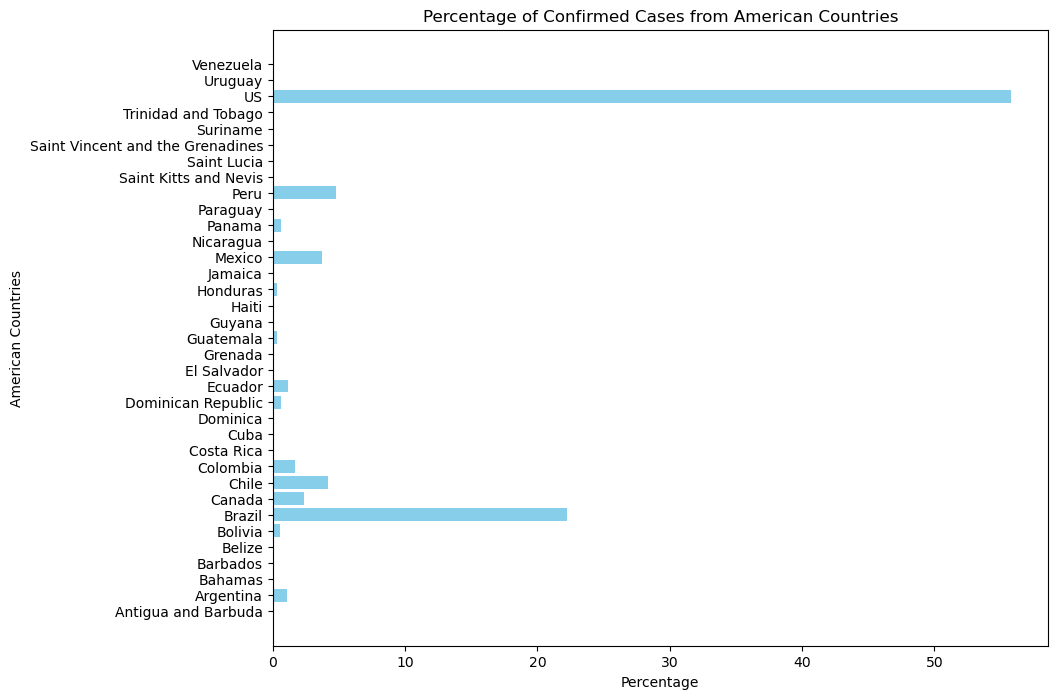
After having our dataset clean and ready to operate, I was in charge in work in the data visualization. I decided to work with bar charts for the visualization and understand better what information the data offered us.

The work was done with the WHO Region column, and different graphs were created for each Covid case. This way, it was possible to visualize that the most affected areas were Europe and America, specifically Russia, Spain, Italy, the United States, and Brazil.

These countries showed the highest percentage in each case studied. A clear example is the case of the United States, where upon analyzing the confirmed cases, it was evident that it had over 50% of the cases. However, when analyzing the recovered cases, it was observed that 35% belonged to this country. This may indicate that it is very likely that this country had the necessary facilities and equipment to care for it is population.

Later, we could observe that in active cases, once again the United States appeared in first place with 70%. This leads us to think that despite the measures and efforts the country was taking to combat this virus, the population may not have been very committed to these implement prevention measures.

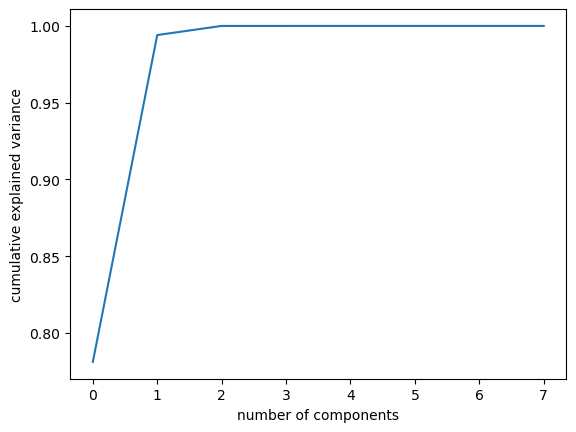
We could also assume that since the United States is a tourist destination, perhaps strict measures were not taken at airports, and the entry of people from other countries was allowed in order to avoid a negative impact on the economy.







As other collaboration that I made in the project was the dimensionality reduction, I applied the PCA with the data not standardized, of the 8 columns that were originally in the data after performing the PCA, 3 columns were obtained.



Also I was in charge to apply machine learning models learned in class, as a team we decided to use apply the Hyperparameter tuning with the models the KNeightbours Regressor and Random Forest Regressor at 10, 20 and 30 % of testing and 90, 80, and 70% of training. The resume of the results obtained are:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Test Split** | | | | | |
|  | **10%** | | **20%** | | **30%** | |
| **Model** | **Kneighbors Regressor** | **Random Forest Regressor** | **Kneighbors Regressor** | **Random Forest Regressor** | **Kneighbors Regressor** | **Random Forest Regressor** |
| **Mean Absolute Error** | 40,62 | 114,49 | 0 | 0 | 0 | 0 |
| **Accuracy** | 0,99 | 0,99 | 1 | 1 | 1 | 1 |

Where we can see that the MAE is higher in both models with 10% of testing and 90% of trainig.

# 16. References

[www.kaggle.com](http://www.kaggle.com). (2020). COVID-19 Dataset. [online] Available at: <https://www.kaggle.com/datasets/imdevskp/corona-virus-report?select=covid_19_clean_complete.csv>.

www.ibm.com. (n.d.). CRISP-DM Help Overview. [online] Available at: https://www.ibm.com/docs/zh/spss-modeler/18.0.0?topic=dm-crisp-help-overview [Accessed 1 Nov. 2023].

Eryk Lewinson (2020) Python for Finance Cookbook : Over 50 Recipes for Applying Modern Python Libraries to Financial Data Analysis. Birmingham, UK: Packt Publishing.

www.spureconomics.com. (2023). Dickey Fuller Test of Stationarity - SPUR ECONOMICS. [online] Available at: <https://spureconomics.com/dickey-fuller-test-of-stationarity/>.

CCT (2023). Harvard Referencing Guide. Available at: <https://cct.libguides.com/ld.php?content_id=34285195> [Accessed 1 Nov. 2023].