May 7, 2020

EXPLORING NUMBEr OF COVID-19 CASES AND BUILDing A MODEL FOR PREDICTING FUTURE Cases

# Contents

[Contents 1](#_Toc39792214)

[1 Introduction 3](#_Toc39792215)

[1.1 Background 3](#_Toc39792216)

[1.2 Problem 3](#_Toc39792217)

[1.3 Interest 3](#_Toc39792218)

[2 Data downloading and formatting 4](#_Toc39792219)

[2.1 Data source 4](#_Toc39792220)

[2.2 Data Cleaning and Formatting 4](#_Toc39792221)

[3 Data Visualization and Exploration 6](#_Toc39792222)

[3.1 Bar Charts 6](#_Toc39792223)

[3.2 Choropleth Map 6](#_Toc39792224)

[4 Modeling and Regression: 7](#_Toc39792225)

[4.1 Results 7](#_Toc39792226)

[5 Discussion and Conclusion 8](#_Toc39792227)

# Introduction

## Background

Novel Coronavirus (COVID-19) is a new strain of coronavirus which may cause illness in animals or humans.  In humans, several coronaviruses are known to cause respiratory infections ranging from the common cold to more severe diseases. First identified in a cluster with pneumonia symptoms in Wuhan city, Hubei province of China.

## Problem

The purpose of this report is to provide exploratory findings and depiction of number of registered cases for Covid-19 pandemic disease up to 07.05.2020.

## Interest

This report will provide an easy to interpret mean of viewing and exploring data, this will be achieved by using heat-maps.

Also, port of this project a Regression model will be built to predict the future amount of infections in the world based on history.

# Data downloading and formatting

## Data source

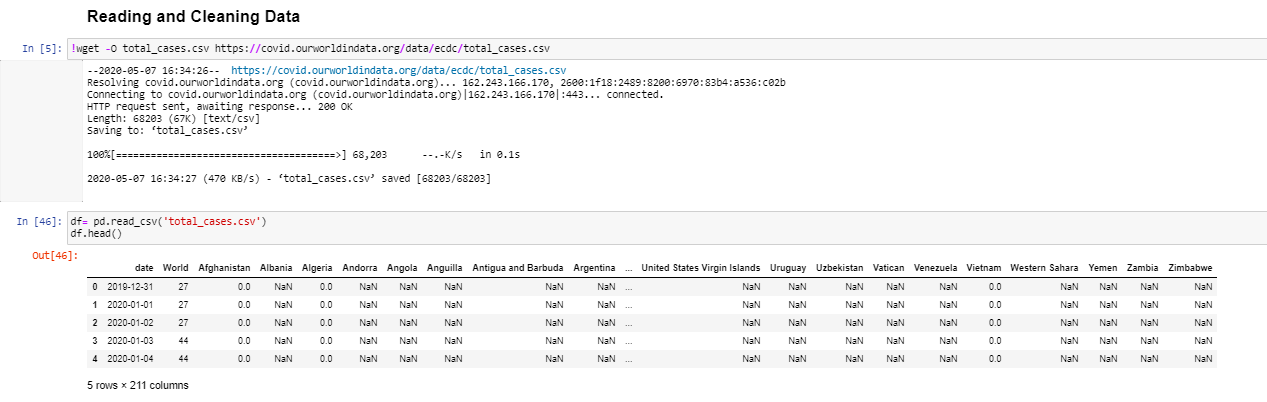
The data set is retrieved from the web location:

<https://covid.ourworldindata.org/data/ecdc/total_cases.csv>

The data properties are as follows:

* Data is a CSV file.
* Data contains details about:
  + Data.
  + Countries in the world
  + Number of cases in the same date in each country
  + Total number if cases in the world

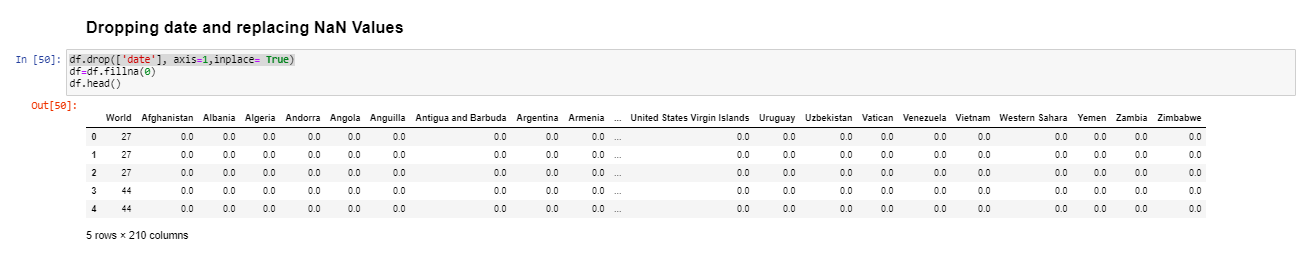




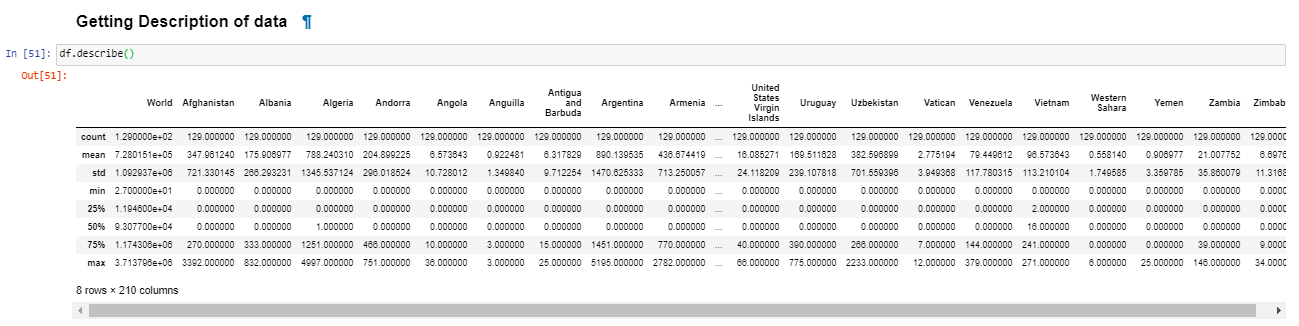
## Data Cleaning and Formatting

The following processes were done to the data to provide better outputs:

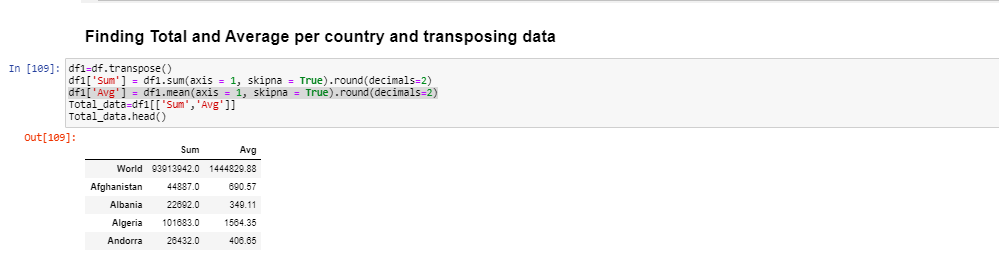
* NaN data are replaced with Zeros “as they represent no registered cases”.
* Date is dropped as we are interested in data order rather than the actual date itself.



* Description about data is found



* Total cases per country is found, and data about top 5 countries is plotted in several ways to provide explanatory information.



* Data is normalized

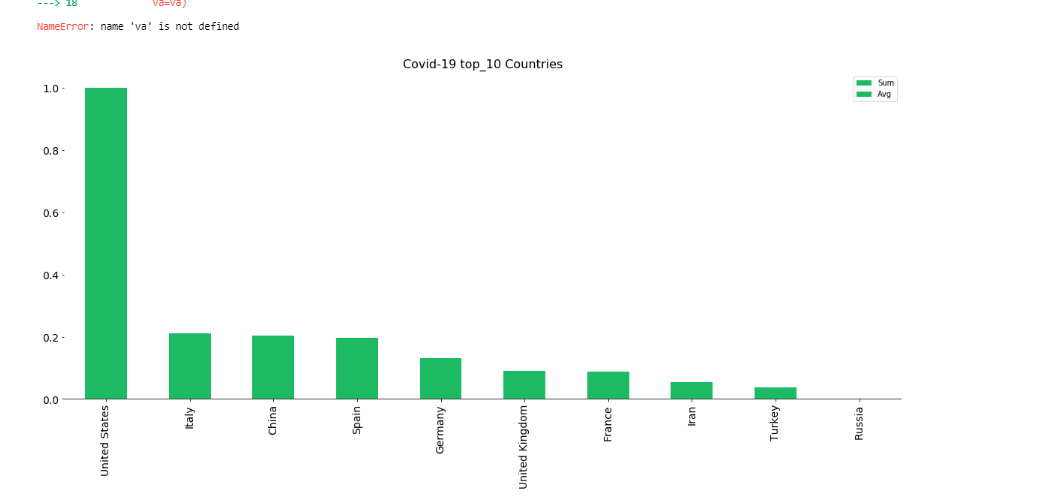


# Data Visualization and Exploration

Two mechanisms were used to visualize data to understand in a better way the spread of Covid-19 cases:

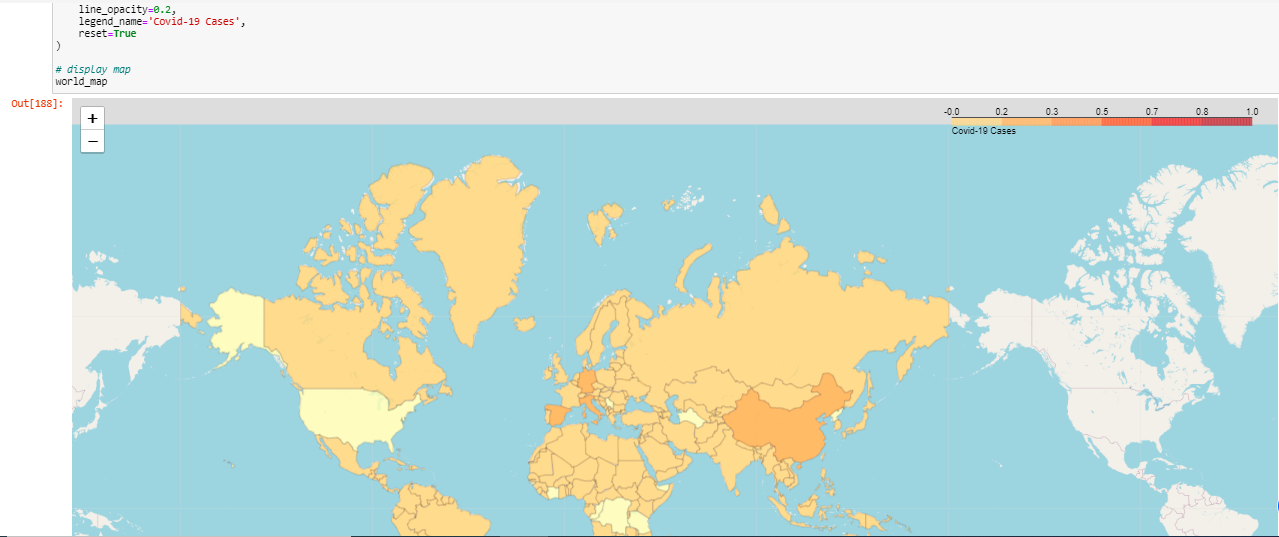
## Bar Charts

Bar chart was used to provide the distribution for the top 10 countries



## Choropleth Map

Choropleth map was used to find the distribution worldwide.



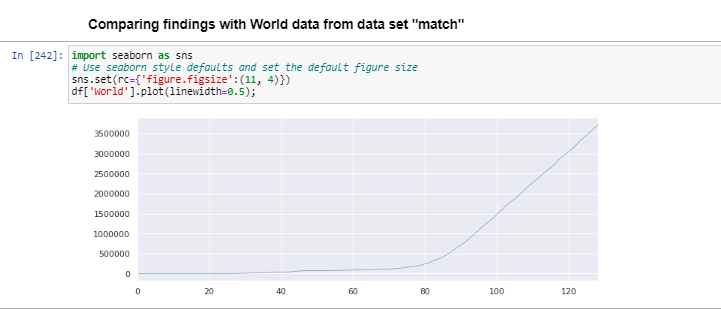
# Modeling and Regression:

(ARMA) Autoregressive–moving-average model was used to build a regression model for predicting number of cases in the world based of the history for the previous days. ARMA is a statistical analysis of time series technique, autoregressive–moving-average (ARMA) models provide regression based on time series.



## Results

Predicted results were compared with the actual results form data set and a good match was found.



# Discussion and Conclusion

Several techniques can be used to provide more readable results when visualizing spread of the Covid-19, for instance Choropleth maps and heat maps can show worldwide, or county based distribution, while bar graphs can show ranking for most affected locations.

Time series can be used to predict the spread of Covid-19 and the places those will be most affected, based on the history provided so far.

Using data science is essential for medical research support, and much more findings can be obtained to support medical researchers.