Logo, company name

Description automatically generated

**Data Analysis Visualization of COVID-19 Dataset**

**Group Name:  DATA ANALYZERS (ADT PROJECT GROUP 5)**

**Student Names: Aaditya Pradipbhai Parekh (110084734)**

**Devanshi Shah (110083753)**

**Mayank Chauhan (110076457)**

**Kshitiz Choudhary (110075501)**

1. **Project Definition :-**

In December 2019, a cluster of infected patients with unknown cause was reported to WHO (World Health Organization) by Wuhan, China. A brand new virus which has not been discovered was named as novel Coronavirus in January 2020. COVID-19 was declared as a pandemic by WHO.   
  
Since the beginning, Covid-19 has had a firm hold on people’s lives and some countries have been struggling with an increase in infections. With the help of data analysis, one can examine how different countries are performing in terms of controlling the outbreak of novel Coronavirus. Predictions are made using the dataset provided to the nation assisting them in determining how much they can control the pandemic or how much they should direct preventive measures.

The main objective of the project is to provide data analysis of Covid-19. Different situations, such as the most affected countries by the virus, have been analyzed through plotting of data. Through this project, a step is taken to help citizens in understanding the spread and visualize cases in their nation. To display and analyze the growth of cases and recovery graph, analysis of data from multiple countries is merged. This project also provides information on how prepared and well-advanced a nation is at controlling the spread. To analyze how much a country is being impacted by and recovering from each day, comparison graphs have also been displayed.

1. **Literature/ Background Study :-**

There are a lot of research papers published that are related to the covid-19. Some of them to name can be research work related to vaccines or other details related to covid-19. Also there are some papers on the medical drugs that can help to recover. People who recovered can shed some light after a deep analysis on how to deal with the active cases. Rajan Gupta and Saibal Kumar Pal, in their research paper 'Trend analysis and forecasting of covid-19 break in India' used exploratory data analysis to report the situation in the time period of January to March in India. They use time series forecasting methods to predict the future trends.

In this paper it is shown that there have been different modelling approaches presented by various researchers for different countries. Models for China [1,2], Italy [3], France [4], India [6] and USA [5]. But all this models are Country wise. The scope of this study is limited to building forecasting models for Indian region and uses time series based forecasting methods which are easy to build and easy to understand in these kind of critical conditions. The study does not include forecasting for any other nation suffering from COVID-19 outbreak [7]. Our purposed model is to do the visualization of covid-19 data world wise.

Another very relevant paper in the journal of 'Chaos, Solitons & Fractals' uses clustering methods to analyze countries on the basis of most affected patients and how they are reacting to it. This paper shows the world scenario first and thereby drilling down to the country of Mexico. Another recently published paper in the journal - 'Chaos Solitons & Fractals' captures the trend of cases and also a prediction using Fourier Decomposition Method. This paper also collects data till 1st week of June and predicts the expected number of cases and deaths in the upcoming days.

Another research paper by Kamlendu Pandey and Ronak Panchal, A Study of Real World Data Visualization of COVID-19 dataset using Python. In this paper the main purpose is to use data-visualisation to communicate clearly and effectively using different graphical representation. In this paper the shown different data visualization libraries. In this paper the researchers identifies following finding based corona virus (COVID-19) datasets - Till 2nd April 2020 Total Cases in India, Date Wise Total Corona cases in India, India's Map with State wise data of Total Cases, Deaths and Cure[8]. In our model we will be using some similar and some different libraries for analysing and producing some meaningful graphical data representation from corona virus dataset.

The scope of this study is limited to just India only and also in this study they are not analysing all the aspect which can be done using the dataset of corona virus. The another limitation of this study is they are doing only the basic analysis of the data but with the dataset it is possible to a deep study of the dataset and will be able to do further analysis of data in different ways. For our model we are going to compare the growth of Covid-19 confirmed, death and recovered cases in the world wise and not just for the major countries that have also been heavily infected. The visualizations are created in Python where using the matplotlib, seaborn, plotly libraries and also date time library for time series data analysis. We will be using the dataset to identifies some different finding like worldwide total confirmed, recovered and deaths. We will be doing the cases density animation on world map. Will be doing the analysis on confirmed cases with choropleth map. Death per 100 and recovered per 100 graphs. Growth rate after the 100 cases and 1000 cases and 10000 cases. We also make a gantt chart which display the first and last case report time. This research paper can also be of help to several other sectors or other branches of healthcare as immunity power is very related to fighting with Covid-19

In our project we also be comparing the severity of COVID-19 with other epidemic outbreaks like EBOLA 2014, and MERS 2012 it will be a benefit to understand the strategies and planning required if we face the same kind of epidemic in the future. The analysis of the data for the COVID-19 will help to understand and fight the pandemic more effectively in the future. The countries like Vietnam, South Korea and Singapore have all experienced previous outbreaks of Middle East respiratory syndrome (MERS) or severe acute respiratory syndrome (SARS) epidemics in the past and have displayed remarkably effective responses to COVID-19, with better control of viral spread and significantly fewer confirmed cases and deaths attributed to COVID-19[9,10].

1. **Proposed model :-**

In this project, the dataset we are using is in CSV format and is updated daily. It is sourced from the <https://github.com/CSSEGISandData/COVID-19>, maintained by Johns Hopkins University Center for Systems Science and Engineering (CSSE) who have been doing a great public service from an early point by collating data from around the world. This dataset is maintained by World Health Organization (WHO) , DXY.cn. Pneumonia. 2020, BNO News, National Health Commission of People’s Republic of China (NHC), Chinc CDC (CCDC), Hong kong Department of Health, Macau Government, Taiwan CDC, US CDC, Government of Canada, Australia Government Department of Health, European Center for Disease Prevention and Control (ECDC), Ministry of Health Singapore (MOH) and Italy Ministry of Health.

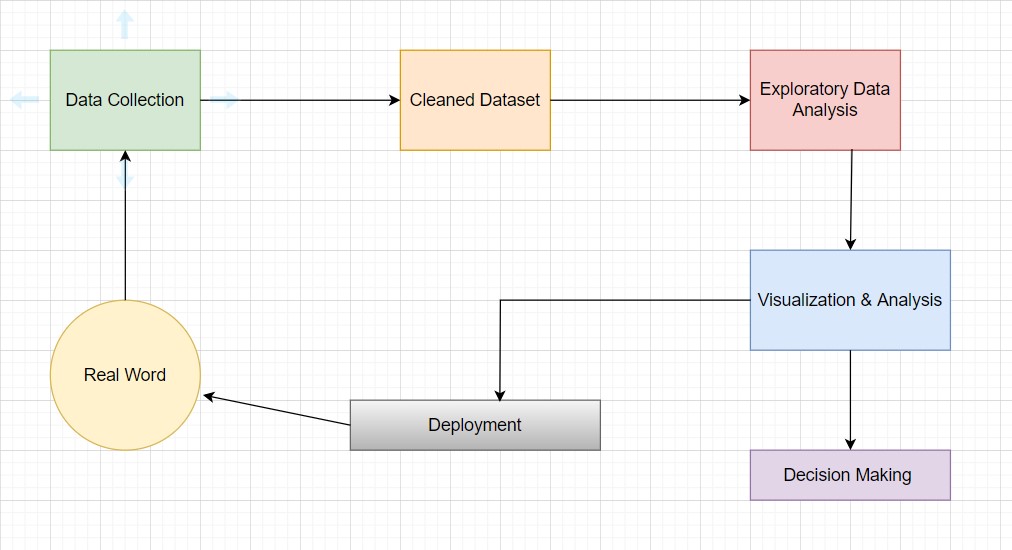


Fig1. Workflow Diagram for Data Analysis and Visualization

In our model we will be taking the dataset from the github directly which will be a cleaned dataset. After

Receiving the dataset we will be doing the Exploratory data analysis and after this step we will be generating various type of graphs which will be done at the visualization and analysis step. Once the graphs is generated and we will receive the data based on the commands given it will be easy to understand and answer various problem statements. Also after the Visualization step the data will be sent for the deployment if any other wise we will get the answers to all the various problem statements.

The main objective of this data visualization is to communicate information simply and efficiently using various graphical representations as it has various conceivable functions in the domain. Visualization is a helpful medium for examining, understanding, and transmitting the information. Python is considered one of the best programming languages for handling data visualization because of its vast and active scientific computing community, as well as its numerous libraries that provide greater flexibility. It may also manage the specific parts of the graphs that are formed and make those requirements code reproducible. Python is also excellent at dealing with data and can manage massive volumes of data without crashing. For data visualization in python, we will use different libraries such as Matplotlib, Pandas, NumPy, Plotly and Folium. Folium makes it easy to visualize data that is being manipulated in python on an interactive leaflet map. It enables both the binding of data to a map for choropleth visualization as well as passing rich vector/raster/HTML visualizations as markers on the map. With the help of the folium library, the world graph shows information in detail about the number of confirmed cases and deaths along with the country and province name. It also has zoom-in and zoom-out features for a magnified view of the map.

By the use data analysis visualization techniques, we have successfully implemented interactive graphs:

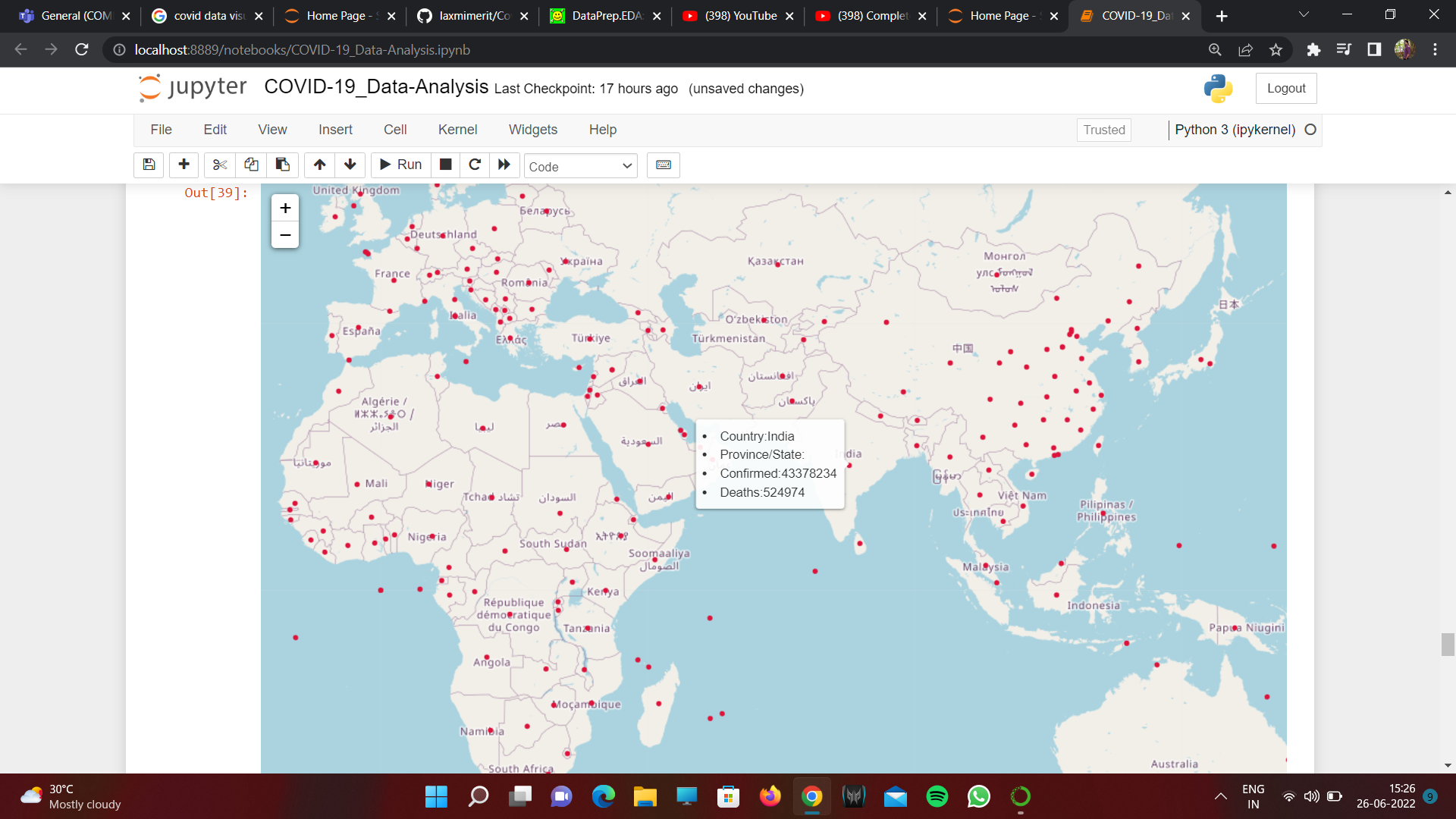
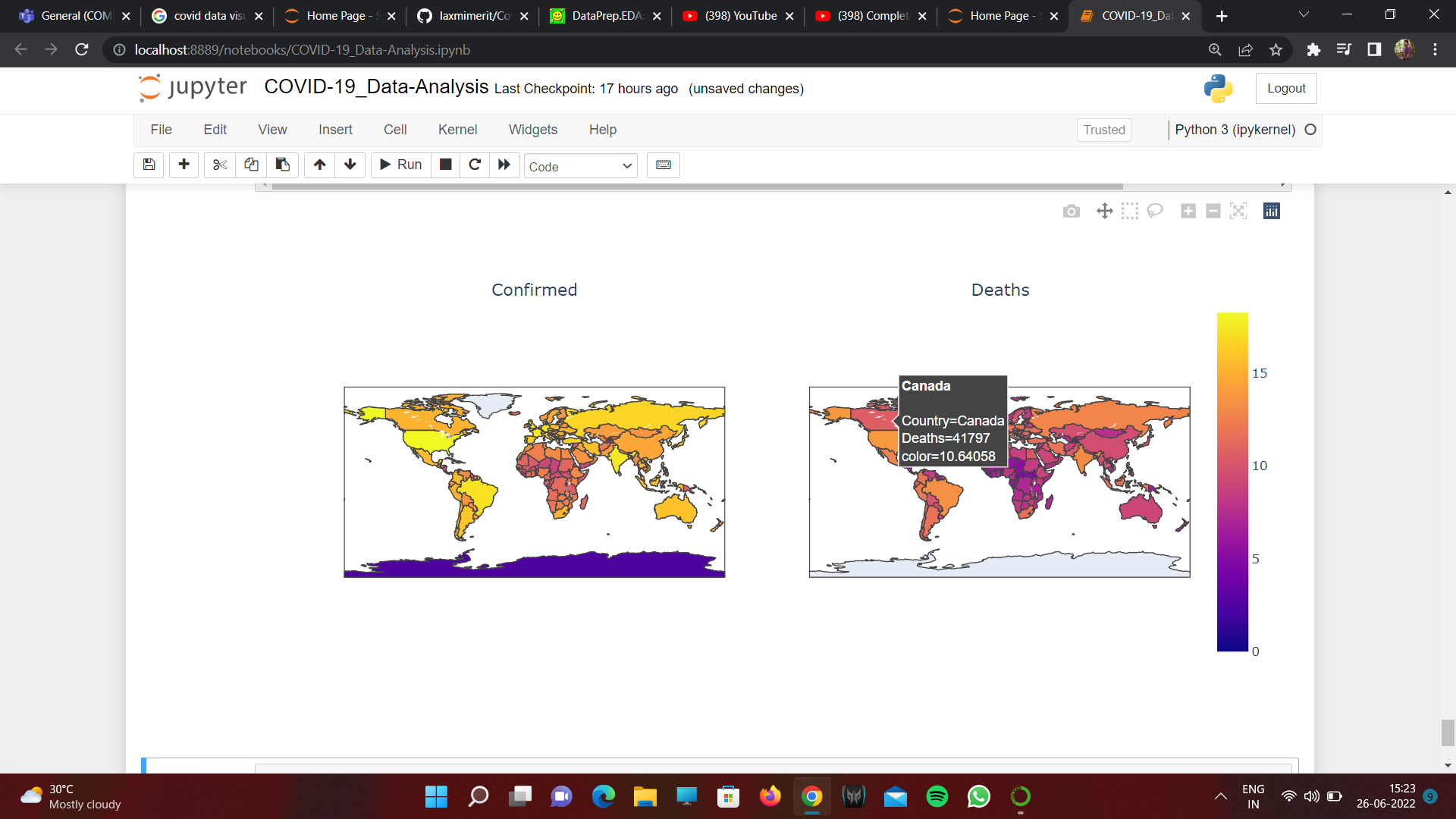
 

Fig2. Worldwide cases on Folium Map Fig3. Confirmed and Death Cases with Static Color Map

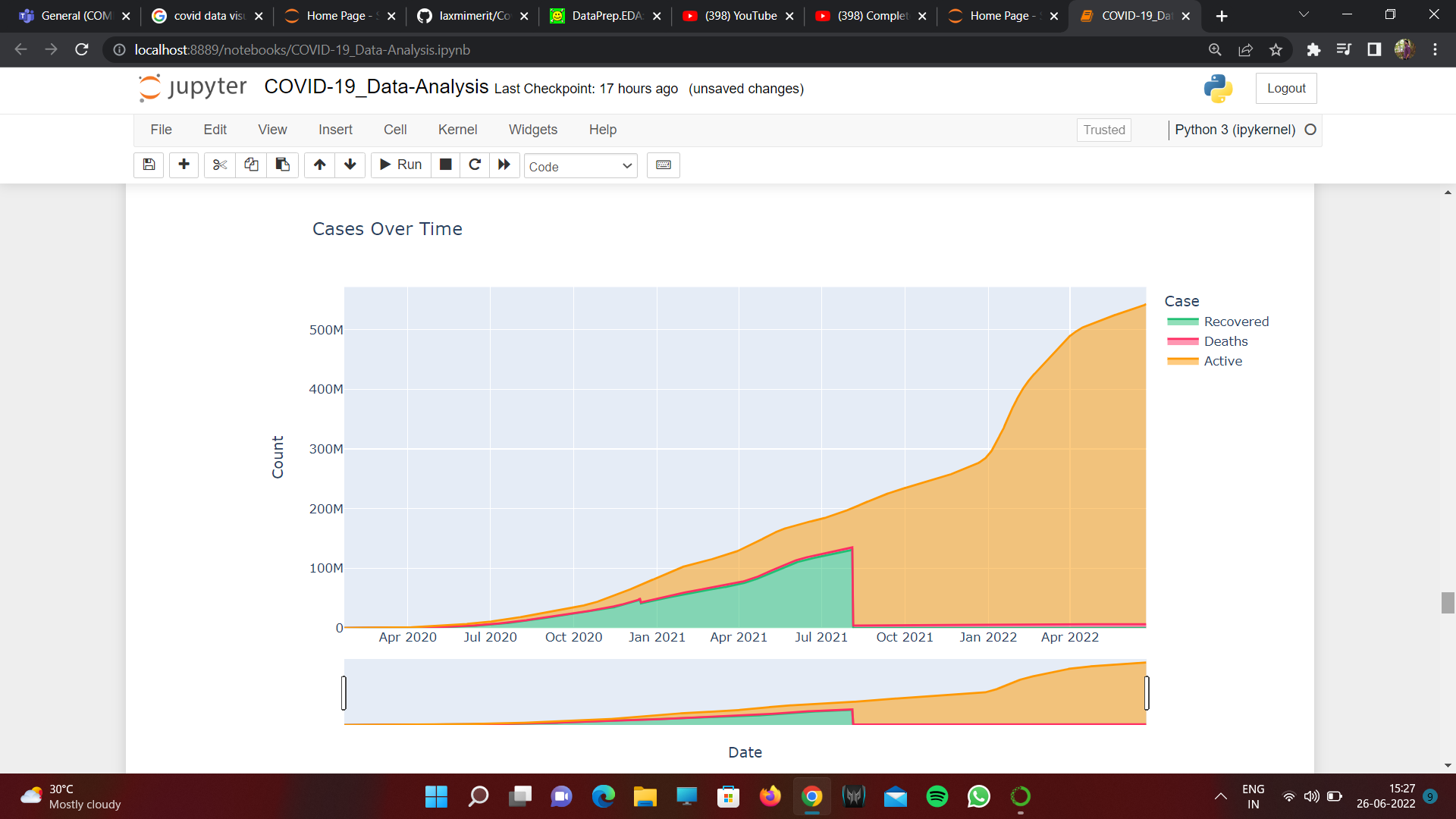
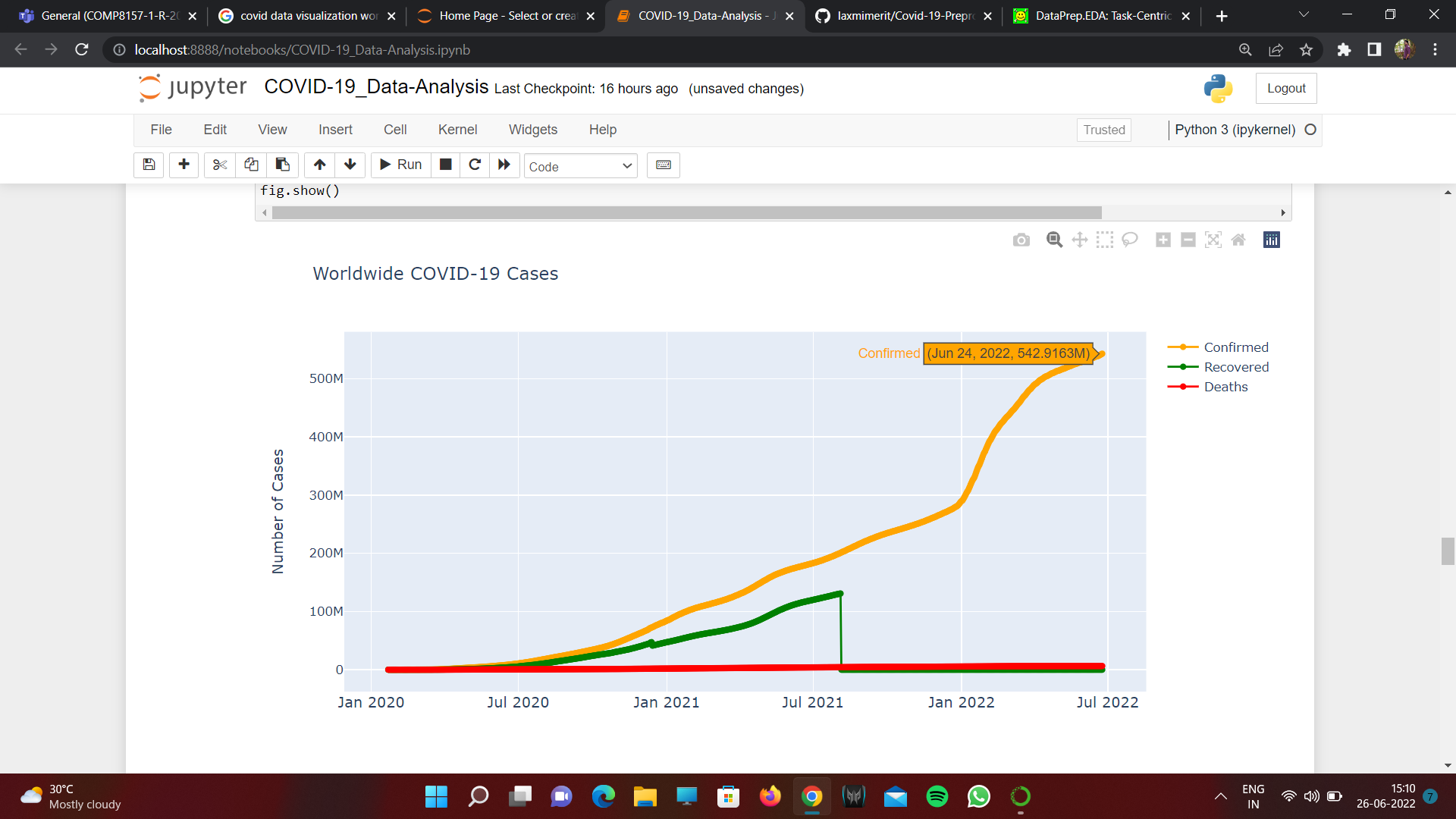
 

Fig4. Recovered, Deaths and Active cases over time Fig5. Worldwide Covid-19 Cases

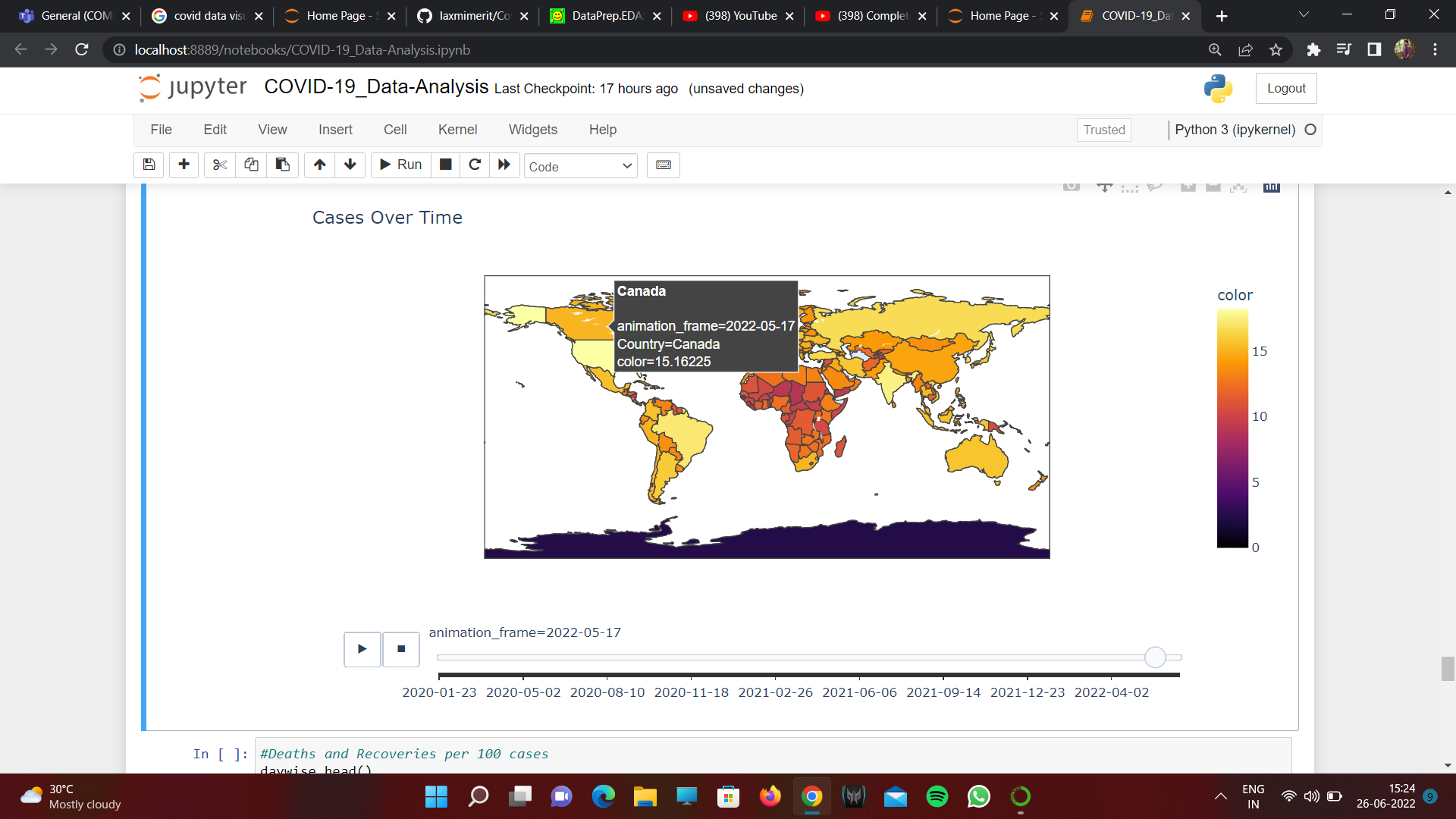
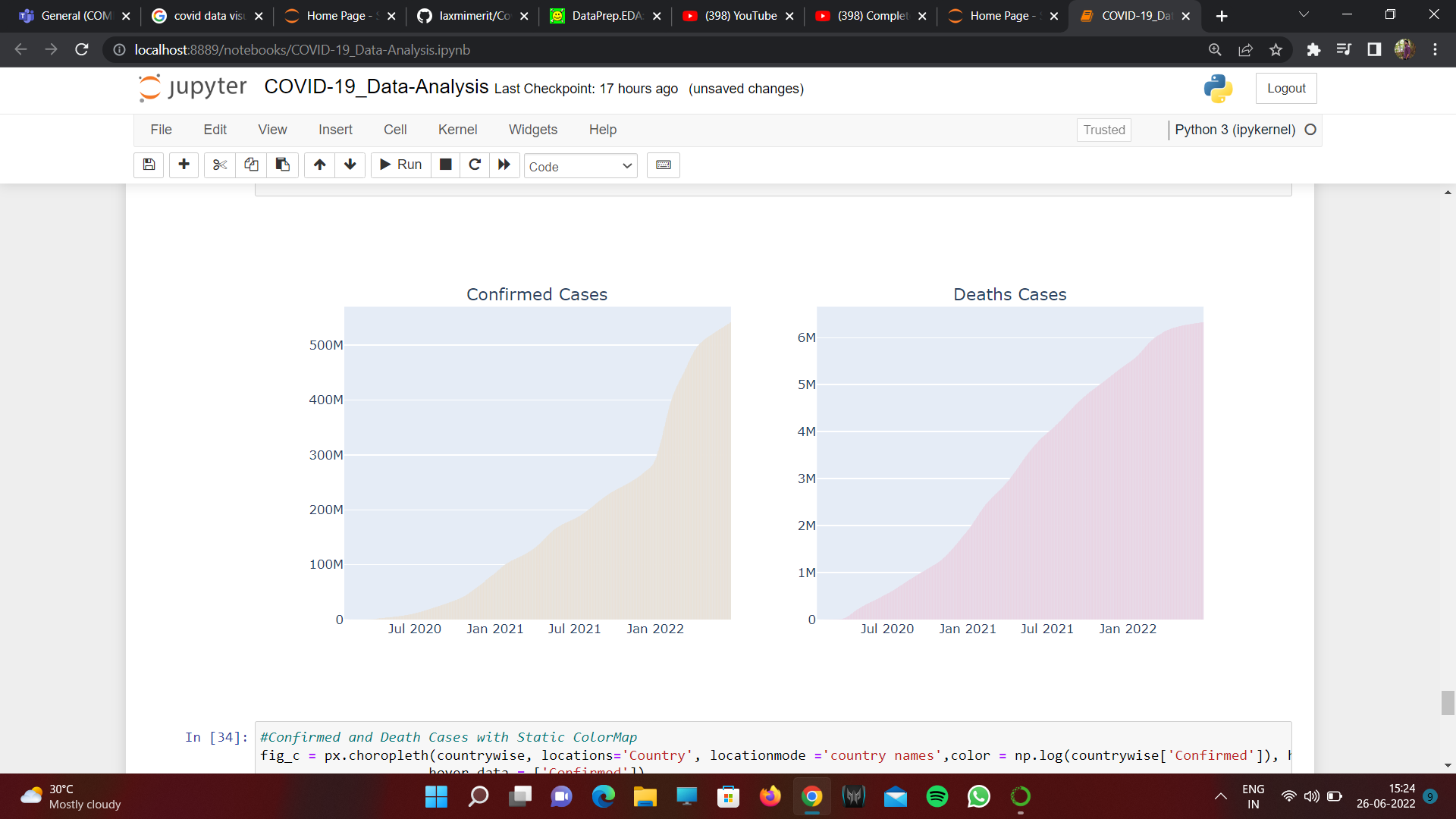
 

Fig6. Confirmed Cases with Choropleth Map Fig7. Confirmed and Death Cases per 100 cases

1. **System Definition (Functional Requirements) :-**

The data visualization uses Folium, plotly, pandas and numpy libraries for representing worldwide total confirmed, recovered and death cases due to COVID-19. The graphs are represented using interactive world maps along with time lapse. Cases density animation using heat maps represents the severity in different parts of the world. The product features/functions that we are planning to implement to accomplish the task are cases over the time with area plot as well as world map, worldwide cases on Folium maps, to display confirmed cases with choropleth maps (cases over time), confirmed and death cases with static ColorMap and deaths as well as recoveries per hundred cases.

The pictorial visualization shows the number of new cases per day in multiple countries throughout the world. The data analysis is done on the top 15 infected countries. It shows confirmed cases, deaths reported, recovered cases, active cases, deaths per hundred cases, recovered per hundred cases, number of new cases, cases per million people and 1 week percentage increase in the cases of these 15 countries. There will be a scatter plot for death versus confirmed cases (cases are on log10 scale). Graphical representations using bar plot and line plot for confirmed, death, new cases, recovered cases for different countries with date. Growth rate of cases is shown after every hundred, thousand, ten-thousand, and hundred-thousand cases.   
  
The representation also shows the tree map analysis of the number of confirmed and death cases, Gantt chart for first and last case report time of different countries, graphical representation of rise of confirmed cases of each country. Also, the visualization displays the comparison of COVID-19 pandemic and other similar epidemics such as SARS, EBOLA, MERS and H1N1. These comparisons are based on confirmed cases, death cases and the mortality rate.

With the help of all these graphical representation of the data we will be able to do the analysis and identify the countries which are required to invest more in the medical infrastructure. Will be able to understand in which countries the spreading of the virus is more and in which it is very minimum. It will also be very helpful in analyzing and understanding the role of the countries which is already faced a pandemic in the past and how they reacted when the covid-19 hit and how also will be able to identify which countries need to take extra steps if any pandemic occurs in the future. As we are using the latest data of the covid-19 we will be able to understand the current scenario of the covid-19 pandemic in some of the countries which is still facing the issue.

1. **References :-**

[1] K. Roosa, Y. Lee, R. Luo, A. Kirpich, R. Rothenberg, J. M. Hyman,… & G. Chowell, “Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020,” Infectious Disease Modelling, 5, 256-263, 2020.

[2] Q. Li, & W. Feng, “Trend and forecasting of the COVID-19 outbreak in China,” arXiv:2002.05866, 2020.

[3] D. Fanelli, & F. Piazza, “Analysis and forecast of COVID-19 spreading in China, Italy and France,” Chaos, Solitons & Fractals,134, 109761, 2020.

[4]C. Massonnaud, J. Roux, & P. Crépey, “COVID-19: Forecasting short term hospital needs in France,” medRxiv, 2020.

[5] P. Liu, P. Beeler, & R. K. Chakrabarty, “COVID-19 Progression Timeline and Effectiveness of Response-to-Spread Interventions across the United States,” medRxiv, 2020.

[6] B. S. Pujari, & S. M. Shekatkar, “Multi-city modeling of epidemics using spatial networks: Application to 2019-nCov (COVID-19) coronavirus in India,” medRxiv, 2020.

[7] R. Gupta, & S. K. Pal, “Trend Analysis and Forecasting of COVID-19 outbreak in India,” medRxiv, 2020.

[8] K. Pandey, & R. Panchal, “A Study of Real World Data Visualization of COVID-19 dataset usingPython,” [International Journal of Management and Humanities](https://www.researchgate.net/journal/International-Journal-of-Management-and-Humanities-2394-0913) 4(8):104-107, April 2020.

[9] Seven things countries have done right in the fight against COVID-19. GAVI, The Vaccine Alliance website; Published June 22, 2020. <https://www.gavi.org/vaccineswork/seven-things-countries-have-done-right-fight-against-covid-19>. Accessed 27 May 2022.

[10] Vietnam: coronavirus pandemic country profile. Our World in Data website; 2020. <https://ourworldindata.org/coronavirus/country/vietnam?country=~VNM>. Accessed 27 May 2022.

[11] Oberoi, A., & Chauhan, R. “Visualizing data using Matplotlib and Seaborn libraries in Python for data science,” International Journal of Scientific and Research Publications (IJSRP), 9(3), p8733, 2019.

[12] Tutorialspoint. (2019a). About the Tutorial Copyright & Disclaimer.Tutorials Point (I) Pvt. Ltd., 1–13. <https://www.tutorialspoint.com/google_colab/google_colab_tutorial.pdf>

[13]. Tutorialspoint. (2019b). About the Tutorial Copyright & Disclaimer. Tutorials Point (I) Pvt. Ltd., 1–13. <https://www.tutorialspoint.com/matplotlib/matplotlib_tutorial.pdf>