JSS SCIENCE AND TECHNOLOGY UNIVERSITY

MYSURU-570006

Department of Information Science and Engineering

Under the Guidance of

Dr. R J Prathibha

Big Data Analytics Synopsis on:

EDA on Coronavirus

Submitted by:

Anirudh D (01JST19PSE001)

Swaroop S (01JST19PSE021)

Introduction

The 2019–20 coronavirus outbreak is an ongoing epidemic of coronavirus disease 2019 (COVID-19), caused by the SARS-CoV-2 virus. There is an obvious concern globally regarding the fact about the emerging coronavirus 2019 novel coronavirus (COVID-19) as a worldwide public health threat. As the outbreak of COVID-19 causes by the severe acute respiratory syndrome coronavirus 2 progresses within China and beyond, rapidly available epidemiological data are needed to guide strategies for situational awareness and intervention. It started its journey in Wuhan city of China and slowly crept into other countries making people helpless. There is no place untouched by this transmittable virus on the earth and now the entire world is under the cruel grip of the lethal virus. Thousands of people died of coronavirus all over the world and uncountable number of people have been suffering from it. New tools of cell and molecular biology have led to increased understanding of intracellular replication and viral cell biology, and the advent in the past five years of reverse genetic approaches to study coronaviruses has made it possible to begin to define the determinants of viral replication, transspecies adaptation, and human disease.

An exploratory data analysis with visualizations has been made to understand the number of different cases reported (confirmed, death, and recovered) in different provinces across the world. The epidemiological data need to be analysed in a way so that the exploratory data analysis (EDA) methods and visualization model will increase the situational awareness among the mass community in upcoming days. Health workers, governments, and the public, therefore, need to cooperate globally to prevent its spread. Overall, at the outset of an outbreak like this, it is highly important to readily provide information to begin the evaluation necessary to understand the risks and begin containment activities.

Literature Survey

Sl no.	Autor Names	Title	Year of Publication & Source	Dataset	Remarks
1	Samrat Kumar Dey, Md. Mahbubur Rahman, Umme Raihan Siddiqi, and Arpita Howlader	Analyzing the Epidemiological Outbreak of COVID-19	Journal of Medical Virology - 2020	COVID 19 Dataset	Data visualization model will be more effective to understand the epidemic outbreak of this severe disease. Visualization model like Map view and Tree Map View provides an interactive interface.
2	Na Zhu, Ph.D., Dingyu Zhang, M.D., Wenling Wang, Ph.D., Xinwang Li.	Coronavirus from Patients with Pneumonia in China	The New England Journal of Medicine-2020	Corona Virus Data WHO	This Paper gives knowledge about the Novel Coronavirus affect in China. The study is done on patients with pneumonia with air way block in them.
3	Alissar Nasser, Denis Hamad, Chaiban Nasar	Visualization Methods for Exploratory Data Analysis	IEEE - 2006	Synthetic Dataset, Wine Dataset and Iris Dataset	Linear and nonlinear projection methods for exploratory data analysis. PCA, KPCA, Sammon, and CCA methods are compared and tested on synthetic and real datasets.
4	Changhui YU	Research of time series data based on exploratory data analysis and representation	IEEE - 2016	Air Quality Data	The daily environment air quality data of Wuhan from Jan 2013 to May 2016. find The tendency and hidden patterns in the data. The daily air quality data is analysed the trend of air pollution.
5	Andrea Batch and Niklas Elmqvis	The Interactive Visualization Gap in Initial Exploratory Data Analysis	IEEE – 2018	Nil	Data scientists and other analytic professionals often use interactive visualization in the dissemination phase at the end of a workflow during which findings are communicated to a wider audience.

Existing System

There is an obvious concern globally regarding the fact about the emerging coronavirus 2019 novel coronavirus as a worldwide public health threat. As the outbreak of COVID-19 causes by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) progresses within China and beyond, rapidly available epidemiological data are needed to guide strategies for situational awareness and intervention. The recent outbreak of pneumonia in Wuhan, China, caused by the SARS-CoV-2 emphasizes the importance of analysing the epidemiological data of this novel virus and predicting their risks of infecting people all around the globe.

Problem Statement

To analyse the epidemiological outbreak of COVID-19. A visual exploratory data analysis approach to track and visualize the spread of the virus. A Visual Exploratory Data Analysis model helps to understand the consequences of the COVID-19 outbreak.

Scope and Objectives of Problem Statement

The Objectives are as follows:

- To understand the dataset attributes and values for analysis of COVID-19 data.
- To clean the Dataset for any noise, out layers and extract only the necessary information.
- To preform exploratory data analysis on the cleaned dataset by making visual reports and understand them which is required to forecast the spread of COVD-19.

Hardware Requirements

- ✓ Processor : Intel I3
- ✓ Speed : 2.5Ghz
- ✓ RAM : 4GB
- ✓ Hard Disk : 500GB

Software Requirements

- ✓ Operating system : Windows 10
- ✓ Coding Language : Python
- ✓ Data Base : CSV, JSON
- ✓ Tool : Jupyter Notebook, Colaboratory

Architecture of Proposed Project

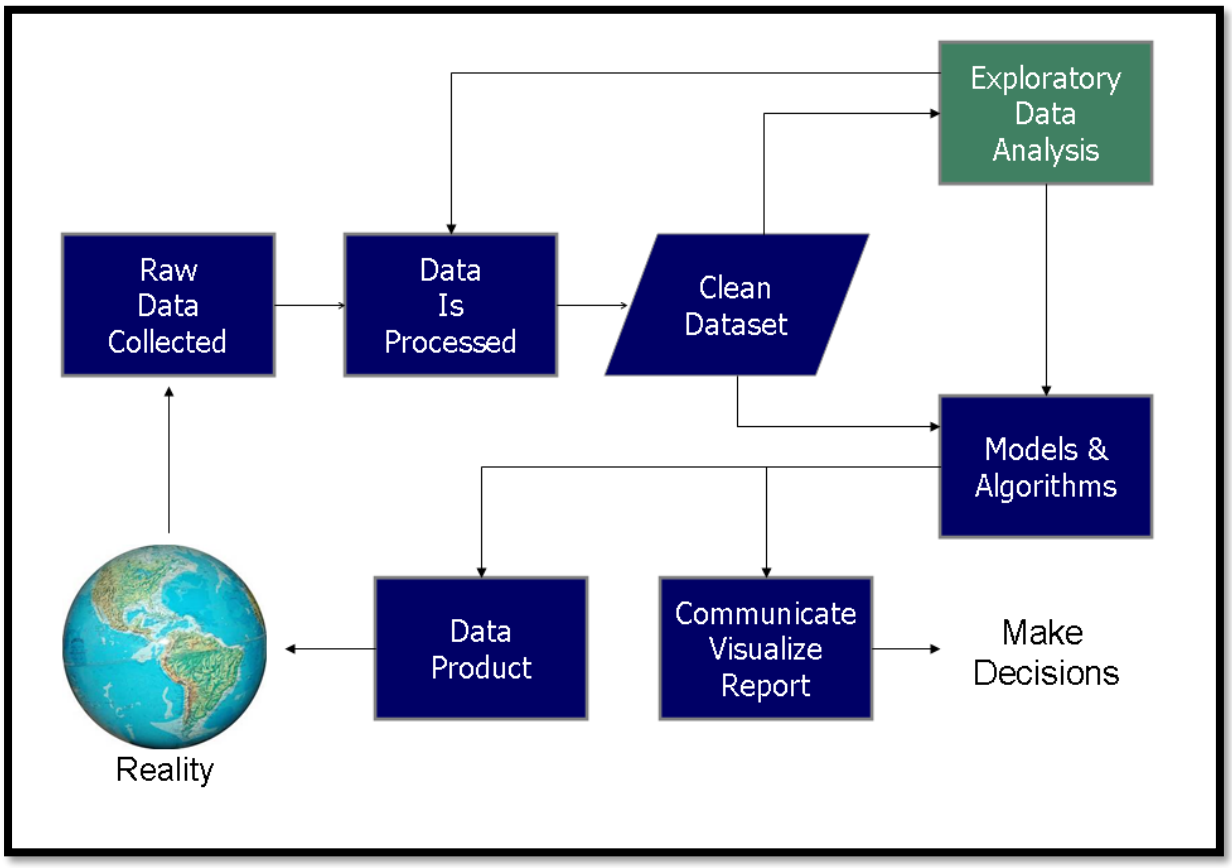

Fig 1. Exploratory Data Analysis

In the Exploratory Data Analysis approach:

1. Data is collected from the environment, represented by the globe.
2. Data is "cleaned" or otherwise processed to produce a data set (typically a data table) usable for processing.
3. Exploratory data analysis and statistical modelling may then be performed.

About the Dataset

COVID-19 Complete Dataset Number of Confirmed, Death and Recovered cases every day across the globe. Dataset is present in WHO website. Data is real time means i.e. it gets updated every 24 hours. The file contains the cumulative count of confirmed, death and recovered cases of COVID-19 from different countries. Each row contains report from each region/location for each day. Each column represents the number of cases reported from each country/region.

Approaches

A visual exploratory data analysis approach is used to analyse the data. It is an approach for analysing datasets to summarize their main characteristics, often with visual methods. EDA is used for seeing what the data can tell us before the modelling task. It is obvious that analysing these data is extremely useful in capturing an epidemic behaviour of this severe disease. We believe this method of analysing data will certainly increase the understanding of the situation and inform interventions.

Evaluation of Proposed System

Exploratory data analysis is generally evaluated in two ways. Non-graphical or graphical approach. In this current project, Graphical approach is being used to analyse. Here, first the ranking of edges is done which is used to identify powerful versus weak relationships. In the second step plotting nodes into more comprehensible scatterplots is used to find patterns and outliers. In the final step we enable algorithms to find well connected communities and make visual report out it. By using this report, we analyse and make decisions. The visualizations have been made in order to understand the number of different cases reported (confirmed, death, recovered) in different parts of the world and also used to forecast the spread of COVID-19 among those parts of the world. So, a user-friendly data visualization model like this will be more effective to understand the epidemic outbreak of this severe disease. Visualization model like Map view provides an interactive interface and visualize each and every raw fact in a comprehensive manner.