2020 Seventh International Conference on Information Technology Trends (ITT). 25-26 November 2020

Data Analysis of Novel Coronavirus Based on Multiple Factors

Harshit Raj
Department of Computer Science
University of Maryland – College Park
College Park, United States
harshitraj0153@gmail.com

Ram Krishn Mishra
Department of Computer Science
BITS Pilani, Dubai Campus
Dubai, United Arab Emirates
rkmishra@dubai.bits-pilani.ac.in

2020 Seventh International Conference on Information Technology Trends (ITT) | 978-1-7281-8379-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/ITT51279.2020.9320887

Abstract – **The novel coronavirus, also known as COVID-19, is the newest strain of coronavirus that causes respiratory infections in humans. This virus is affecting the entire world and has impacted many important sectors, such as travel, economy, education, and hospitality. Many researchers are trying to find effective methods in an attempt to limit the spread of the virus and find a cure for infected people. The study assesses the impact of the virus in 212 countries using advanced data visualization techniques. The research analyzes multiple factors impacting the spread of the virus; Population, Population Density, Median Age, Human Development Index (HDI), Number of COVID-19 Cases, and Number of COVID-19 Deaths. The research incorporates R-Values and Linear Regression to highlight trends and patterns observed.**

Keywords – **Data Analysis, Novel Coronavirus, R-Values, Linear Regression, Human Development Index (HDI), Median Age**

I. INTRODUCTION

COVID-19 is the official name of the coronavirus strain, which was identified in humans in December 2019. This new virus is the seventh known coronavirus to humans and originated in the Wuhan, a city in China's Hubei Province with a population of 11 million [1, 2]. As of September 17th, the virus has now spread in more than 200 countries, infecting more than 30 million people and causing over 945,000 deaths.

COVID-19 virus causes respiratory illness and is extremely contagious. The virus can spread through the air; if an infected person sneezes or coughs close to someone or touches infected things, there is a high probability of catching the virus. Anyone can be infected by this virus, irrespective of age or gender; however, the fatality rate is high for people with existing medical conditions and people above the age of 60 years [3]. The symptoms vary; some patients are asymptomatic; others experience flu-like symptoms, while some get so sick that they need to be hospitalized. These mysteries of COVID-19, which was declared a global pandemic in March 2020, have sent medical experts racing to uncover its secrets and find a remedy [4].

The impact of this pandemic is unprecedented. Apart from health, in order to prevent the spared, various countries went into lock-down, restricting activities and movement, destroying the economy [5]. Multiple countries are experiencing the most significant economic contraction is recorded history, effecting jobs: loss of employment, livelihood, education: education system, social life-impacting the mental health of various people, law and order, and the virus is still not under control. Various studies are being conducted to help control the spread and bring back normalcy [6].

II. BACKGROUND

A study from March 2020 by Jingyuan Wang and co researched and analyzed the relationship between the spread of COVID-19 and temperature & humidity. The investigation looked at the novel coronavirus cases data from 100 Chinese cities from January 19 to February 10, 2020. Furthermore, the data for COVID - 19 cases for 1,005 counties in the United States from March 15 to April 25, 2020. Statistical analysis was conducted in order to determine the relationship between transmissibility of the novel coronavirus and the temperature & humidity. The study concluded that high temperatures and high humidity reduce the transmission of the virus [7].

Another study by Joacim Rocklö and Henrik Sjödin from Umeå University explored the relationship between population density and the spread of COVID-19. The study concluded that the recommendation by the WHO of the one-meter distance between people coughing and sneezing is more challenging in high population densities. The study suggested that the greater the population density, the greater the spread and transmission of COVID-19 [8].

A paper by Denes K.A.Rosario and co investigated the relationship between the novel coronavirus and the weather. The study aimed to evaluate the correlation between the weather, humidity, rainfall, wind speed, and solar radiation with COVD-19. The study was conducted in a tropical country, Brazil (Rio de Janeiro). The study concluded; all the factors discussed in the paper showed a negative correlation with the novel coronavirus. Therefore, hotter tropical climates are a factor that can suppress the spread of the novel coronavirus to some extent [9].

These days, machine learning has picked up significance in prediction by learning the experience from the data. It is regularly utilized in the field of data analytics, where a dynamic model is built to achieve some productive outcomes for prescient examination [10]. This study aims to understand and analyze different factors that result in the spread of the novel coronavirus. The new aspect that will be introduced in this paper will be the relationship between the spread of COVID-19 and the Human Development Index (HDI) of a country. The HDI accesses three dimensions of human development: education, life, and per capita income. The HDI is the geometric average normalized indices for each of the three indicators. This study will also take into account median age, population, and population density as factors when conducting the analysis.

III. METHODOLOGY

This study is being conducted to discover the correlation between different factors in relation to coronavirus. Figure 1 shows the process required to find accurate correlations. The process starts with data collection from the available and accessible sources, followed by preprocessing data and selecting the attributes. Visualization methods can then be used to evaluate and analyze the data. Finally, the prediction line is used to find trends and correlations.

A. Raw Data Collection

This is the first step in gathering needed data from available and accessible sources. The identification of the appropriate data source is one of the most challenging steps. In this research, the data is obtained from multiple sources, including Kaggle, United Nations Development Program, and Worldometer [12, 13, 14]. The data set consists of attributes such as population, population density, median age, Human Development Index (HDI), the total number of coronavirus cases, and the total number of deaths due to the virus for each country.

B. Raw Data Collection

This is the first step in gathering needed data from available and accessible sources. The identification of the appropriate data source is one of the most challenging steps. In this research, the data is obtained from multiple sources, including Kaggle, United Nations Development Program, and Worldometer. The data set consists of attributes such as population, population density, median age, Human Development Index (HDI), the total number of coronavirus cases, and the total number of deaths due to the virus for each country.

C. Data Processing

Data processing requires data cleaning and arranging them in a correct format to enable validation of data and its fitness for the research. From the original dataset, the data is transformed, which can be used for the next phase. The null values are removed for comparisons using algorithms written in Phyton3.

D. Data Analysis

Once data processing is done, the raw data needs to be analyzed using various visualization techniques. This Data analysis phase uses various machine learning algorithms to understand trends and patterns in the data.

E. Calculation for Perdiction Line

The relationship between two variables is modeled through Simple linear regression, wherein it can be precisely estimated how much Y will change when X changes by a specific amount. With regression, one tries to establish a linear relationship between the two chosen variables.

$$\boldsymbol{Y = b_0 + b_1 X} \qquad \textbf{(1)}$$

WHEREIN:

Y: Target
b_0: Constant or Intercept
b_1: Slope
X: Feature Vector
b_0 and b_1 both appear in R output as coefficients.

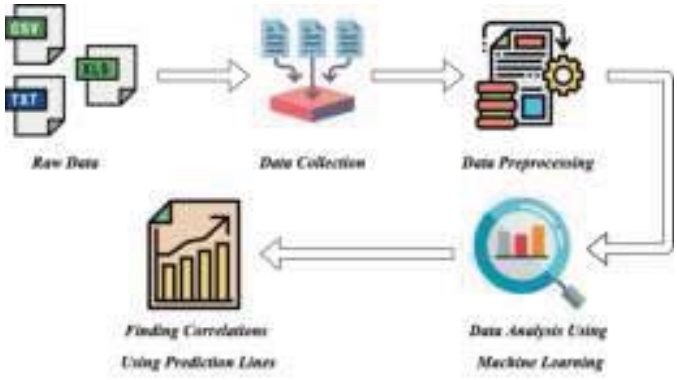

Figure 1: Methodology to Find Correlations

IV. EXPERIMENTAL RESULTS AND ANALYSIS

Python3 Libraries were used to perform analysis of novel coronavirus. This section shows the results of analyzing the dataset with treads through the line of best fit.

a) ***Heat Map and R Values:*** Figure 2 is a Heat Map illustrating correlations for different attributes. Table I is a table showing the R value (ranged between -1 and 1) showing positive and negative correlations.

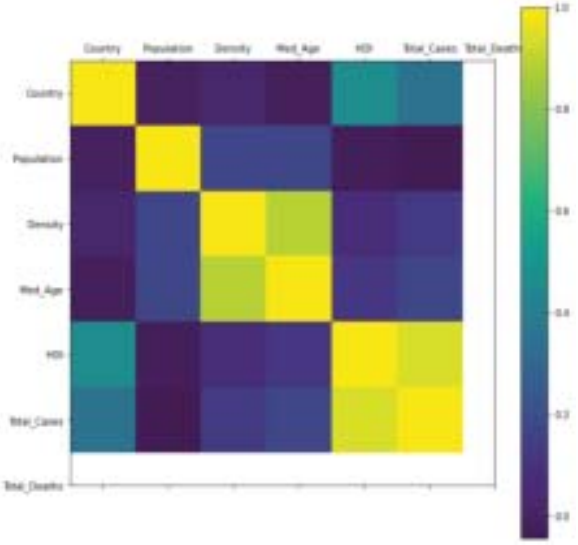

Figure 2: Heat Map for Different Attributes

TABLE I. R VALUES FOR DIFFERENT ATTRIBUTES

	Population	Population Density	Median Age	HDI	Total Deaths	Total Deaths
Population	1.000000	-0.012050	0.017086	-0.019669	0.309609	0.243520
Population Density	-0.012050	1.000000	0.166575	0.170743	-0.032713	-0.041771
Median Age	0.017086	0.166575	1.000000	0.888248	0.135681	0.185752
HDI	-0.019669	0.170743	0.888248	1.000000	0.175709	0.217283
Total Deaths	0.309609	-0.032713	0.135681	0.175709	1.000000	0.930552
Total Deaths	0.243520	-0.041771	0.185752	0.217283	0.930552	1.000000

b) ***Human Development Index versus Total Number of Coronavirus Cases:*** *Figure 3 is a graph that shows the relationship between the Human Development Index (HDI) of a country and the total number of COVID-19 cases in the country. From the line of best fit, there is an evident upward trend suggesting as HDI increases the number of COVID-19 cases increases.*

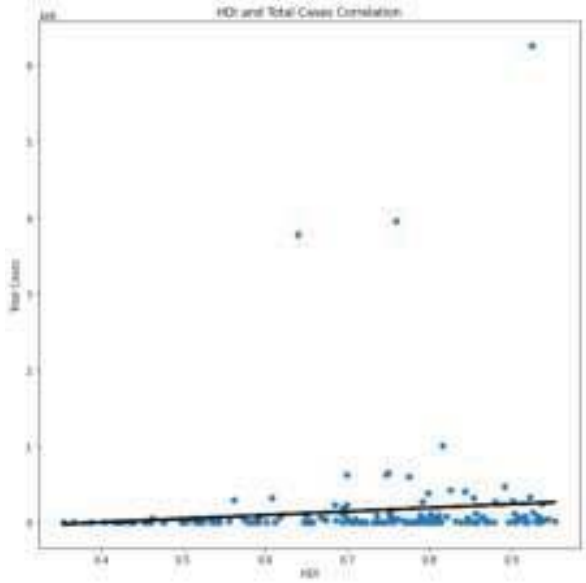

Figure 3: Plot Between HDI and Total Cases Correlation

c) ***Human Development Index vers Deaths due to Coronavirus:*** Figure 4 is a graph that shows the relationship between the Human Development Index (HDI) of a country and the number of deaths due to COVID-19 in the country. It is clear, using the prediction line, that there is a positive trend. As the HDI increases the number of COVID-19 deaths increases.

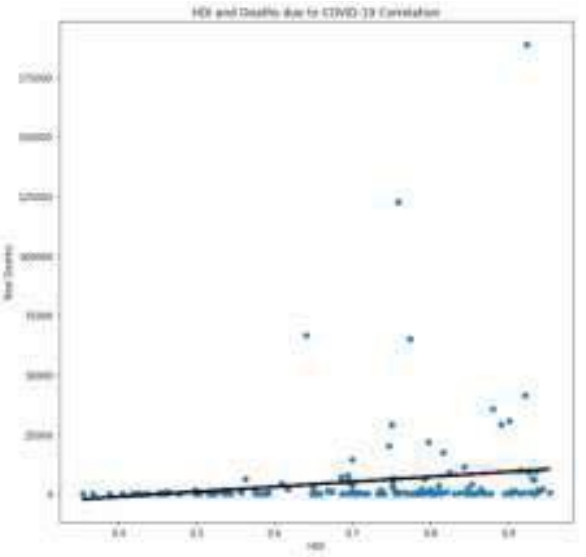

Figure 4: Plot Between HDI and Total Deaths Due to COVID-19 Correlation

d) ***Median Age versus Total Number of Coronavirus Cases:*** Figure 5 is a graph that illustrates the relationship between the median age of a country and the total number of COVID-19 cases in the country. From the prediction line, there is a clear positive trend suggesting as the median age increases the number of COVID-19 cases increases.

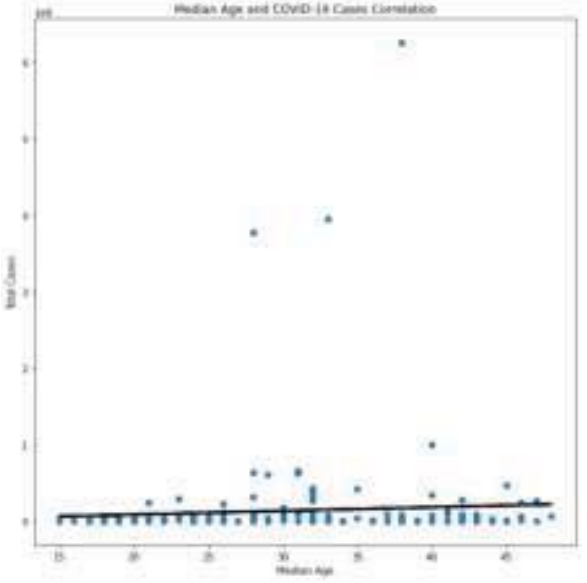

Figure 5: Plot Between Median Age and Total Cases Correlation

e) ***Median Age versus Deaths due to Coronavirus:*** Figure 6 is a graph that shows the relationship between the median age of a country and the number of deaths due to COVID-19 in the country. It is evident using the line of best fit that there is a positive trend. As the median age increases the number of COVID-19 deaths increases.

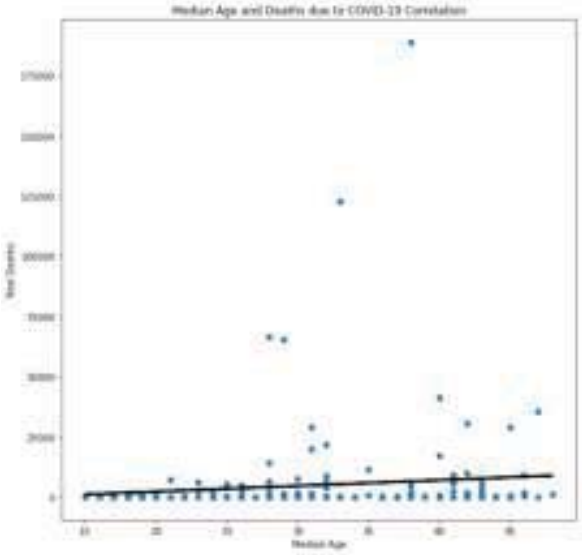

Figure 6: Plot Between Median Age and Total Deaths Due to COVID-19 Correlation

f) ***Human Development Index versus Median Age:*** Figure 7 is that graph that illustrates the relationship between the Human Development Index (HDI) of a Country and the median age of the country. Using the prediction line, there is very distinct positive trend signifying as HDI increases the median age increases.

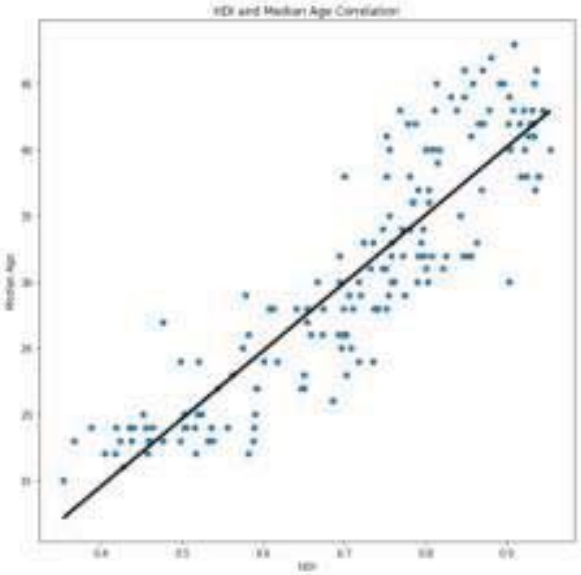

Figure 7: Plot Between HDI and Median Age Correlation

g) ***Top five Countries with the Highest Number of Coronavirus Cases:*** Figure 8 shows the United States has the highest number of COVID-19 cases, followed by Brazil.

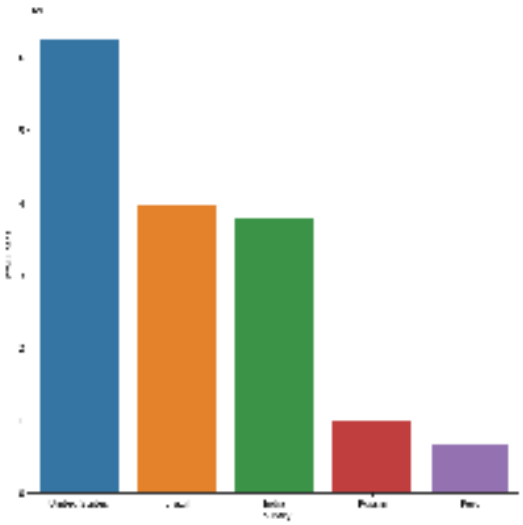

Figure 8: Plot of Top Five Countries with Most COVID-19 Cases

h) ***Top five Countries with the Highest Number of Coronavirus Deaths:*** *Figure 9 shows the United States has the highest number of COVID-19 deaths, followed by Brazil.*

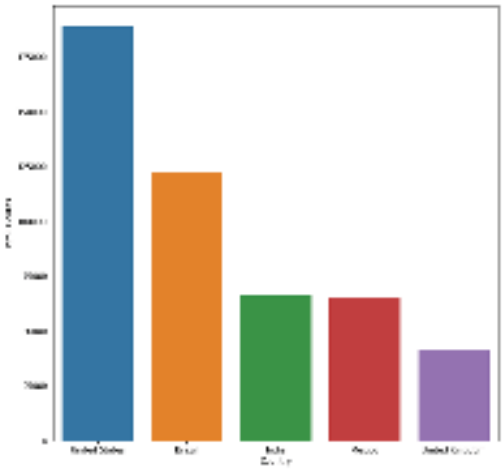

Figure 9: Plot of Top Five Countries with Most COVID-19 Deaths

V. Conclusion and Future Scope

The novel coronavirus is a new virus that is threatening the human race. It is essential to understand the spread and transmission of this virus, as the spread is unprecedented and is infecting everyone who gets in contact with the virus. According to an article by Char Leung from the Deakin University, elderly patients are at a higher risk of mortality [11].

The research conducted suggests that countries with high HDI tend to have a higher number of COVID-19 cases and deaths as compared to countries with low HDI. Furthermore, More Economically Developed Countries (MEDC) have high HDI compared to countries with low HDI because of better per capita income, education, and life expectancy.

People from HDI countries tend to have more disposal income; thereby, they travel more within and outside the country, making them more susceptible to the viruses. Also, an essential factor that needs to be taken into consideration is more testing is happening in countries that have a higher HDI like the United States. This explains the high number of reported COVID-19 cases in high HDI countries.

Furthermore, countries with a greater median age also tend to have more cases and deaths related to the virus. It is evident from the HDI and Median Age graph that there is a very strong positive correlation. This might be because, in countries like the United States and the United Kingdom, people have higher life expectancy due to better healthcare and hygiene, resulting in a greater number of elders as compared to countries with low HDI. As countries that with higher HDI have more elderly people, there are more moralities related to COVID-19.

The research can be extended with more efficient algorithms to identify trends and patterns which can be used for predictive modeling to minimize the impact of future unforeseen pandemics. In addition, more factors can be used to analyze more hidden trends.

REFERENCES

[1] Li JO, Lam DSC, Chen Y, et al, "Novel Coronavirus disease 2019 (COVID-19): The importance of recognising possible early ocular manifestation and using protective eyewear", British Journal of Ophthalmology 2020;104:297-298.

[2] N. Rajapakse and D. Dixit, "Human and novel coronavirus infections in children: a review," *Paediatr. Int. Child Health*, vol. 00, no. 00, pp. 1–20, 2020, doi: 10.1080/20469047.2020.1781356.

[3] Nishiura H, Kobayashi T, Miyama T, et al. Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19). Int J Infect Dis. 2020;94:154-155. doi:10.1016/j.ijid.2020.03.020

[4] Chih-Cheng Lai, Yen Hung Liu, Cheng-Yi Wang, Ya-Hui Wang, Shun-Chung Hsueh, Muh-Yen Yen, Wen-Chien Ko, Po-Ren Hsueh ,"Asymptomatic carrier state, acute respiratory disease, and pneumonia due to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2): Facts and myths" ,Journal of Microbiology, Immunology and Infection, Volume 53, Issue 3,2020, Pages 404-412, ISSN 1684-1182, https://doi.org/10.1016/j.jmii.2020.02.012.

[5] O. Evans, "Socio-economic impacts of novel coronavirus: The policy solutions," *BizEcons Q.*, vol. 7, no. March, pp. 3–12, 2020, [Online]. Available: http://www.bequarterly.rysearch.com

[6] Imran Ali, Omar M.L. Alharbi, "COVID-19: Disease, management, treatment, and social impact", Science of The Total Environment, Volume 728, 2020, 138861, ISSN 0048-9697, https://doi.org/10.1016/j.scitotenv.2020.138861.

[7] J. Wang, K. Tang, K. Feng, and W. Lv, "High Temperature and High Humidity Reduce the Transmission of COVID-19," *SSRN Electron. J.*, 2020, doi: 10.2139/ssrn.3551767.

[8] J. Rocklöv and H. Sjödin, "High population densities catalyse the spread of COVID-19," *J. Travel Med.*, vol. 27, no. 3, pp. 1–2, 2020, doi: 10.1093/jtm/taaa038.

[9] D. K. A. Rosario, Y. S. Mutz, P. C. Bernardes, and C. A. Conte-Junior, "Relationship between COVID-19 and weather: Case study in a tropical country," *Int. J. Hyg. Environ. Health*, vol. 229, no. June, p. 113587, 2020, doi: 10.1016/j.ijheh.2020.113587.

[10] F. A. Binti Hamzah *et al.*, "CoronaTracker: World-wide Covid-19 outbreak data analysis and prediction," *Bull. World Health Organ.*, no. March, p. Submitted, 2020.

[11] C. Leung, "Risk factors for predicting mortality in elderly patients with COVID-19: A review of clinical data in China," *Mech. Ageing Dev.*, vol. 188, no. April, p. 111255, 2020, doi: 10.1016/j.mad.2020.111255.