APPLICATION OF DATABASE SYSTEMS AND ANALYTICS TO COVID-19 DISEASE

DATABASE AND ANALYTICS PROGRAMMING (H9DAP)
PROJECT REPORT

Rahul Kumar x19198400

Diveyesh Arora x18185975

Oluwatobi Ekundayo x19173105

Pragathi Guntamadugu Sudhakar x19179391

MSc. Data Analytics, National College of Ireland

Abstract - **The outbreak of the COVID19 disease created so much pandemonium around the world. This pandemic brought several business activities to a halt, caused the government to spend their emergency fund to keep the people of the nation safe. The Auto ARIMA model used for forecasts, predicted that the number confirmed coronavirus cases and the number of deaths will increase linearly month after month. However, the recovered cases are also predicted to increase linearly. The increase in confirmed cases will possibly cause the medical practitioners figure out some temporary remedy or cure, which will yield an increase in recovered cases. Due to forecast of increase in the spread of coronavirus cases, this paper further emphasises on the need take precautionary measures such as the use of face mask, practise "social distancing" and abide by the preventive protocols laid out by World Health Organization (WHO) and government of each countries of the world.**

Keywords: **COVID19, Forecast, Python, Database, Visualization**

I. INTRODUCTION

The outbreak of the novel coronavirus pneumonia (COVID-19) which was first discovered and reported in Wuhan city, Hubei province, China in December 2019 [1], created so much pandemonium around the world with its rapid spread. This pandemic has brought several business activities to a halt and caused several countries to go into an emergency lockdown to contain the spread of the virus. As of 22 April 2020, death toll across the world has reached 120,000+ deaths. COVID19 has been identified as been consistent with person-to-person transmission of this novel coronavirus in hospital and family settings, and the reports of infected travellers in other geographical regions [2].

The main objective of this project is to analyse the spread of this virus and make some forecast to help health agency strategize better on how make provision for new cases. To achieve this, we obtained related datasets to coronavirus, carried out analysis and predictions using python, stored both gathered data and processed data in separate databases such that referencing or query of any information at any time from the stored data is possible.

The infographics created in this project will be useful to the health sector as well as individuals in search of forecasts and insights about the virus.

- **Research Questions**

Based on the acquired datasets, some specific research questions will be answered for recommendation to the World Health Organisation (WHO).

Q1: Does the forecast model show an increase or decrease in coronavirus cases?

Q2: Which country requires emergency control of the spread of the virus and reduction in death cases?

II. RELATED WORKS

In January 2020, China gave their gathered data to World Health Organization (WHO) for research purposes. World health published this data to the world for research in February 2020. John Hopkins University conducted a research based on the data and measure the symptoms of COVID19 patients [3]. Their research showed that the virus rapidly spread out in 40 provinces of China and by mid-February it had spread across 70 countries, affecting more than 70,000 people around the world. Also, in Mid-February, 900 people died in China. In each country, the numbers have increased rapidly. They also predicted that death toll by 10th February will increase to 2700 in China (lower bound). Chinese government however lockdown the Hubei province in China and were able reduce the number of cases and death in China to a fatality rate of 0.15% of total population by the end of February 2020.

Coronaviruses are large enveloped viruses that are capable of infecting humans and a variety of animals causing mainly respiratory and enteric diseases [4]. COVID-19 is however quite like the two types of novel coronavirus experienced years back.

Middle East respiratory syndrome coronavirus (MERS-CoV) a member of the beta group of coronaviruses, is a zoonotic virus that is transmitted from animals with the virus to humans [5]. A total of 193 deaths and 635 confirmed cases of MERS-CoV infection was reported globally [6]. Transmission of MERS-COV infection in humans was discovered to be a direct contact with the saliva of infected camels or consumption of contaminated milk or meat [7]. MERS-COV was epidemic in Saudi Arabia and the frequency of the cases among patients of

Saudi Arabia were higher in men than women (for those between ages 21 to 60). The infection in Saudi Arabian patients was 62.6% which was almost double of the infection in non-Saudi Arabian patients (37.4%). Confirmed cases of male patients (61.1%) surpassed those of female patients (38.9%). The methodology they applied was the descriptive and comparative statistics using non-parametric binomial test and Chi-square test to describe mortality rate, demographic characteristics, and clinical manifestations [5]. Also, majority of the patients were aged 21–40 years (37.4%) or 41–60 years (35.8%).

Severe acute respiratory syndrome (SARS) is a viral respiratory illness caused by a coronavirus called SARS-associated coronavirus (SARS-CoV) and it is characterized by fever and respiratory symptoms such as cough and shortness of breath [8]. The first SARS case appeared in Southern China (Guangdong Province) in November 2002, after which it was then recognized as a global threat in March 2003. When the outbreak was over and contained in July 2003, about 8,096 cases, including 774 deaths were reported from 29 countries across North America, South America, Europe, and Asia [9].

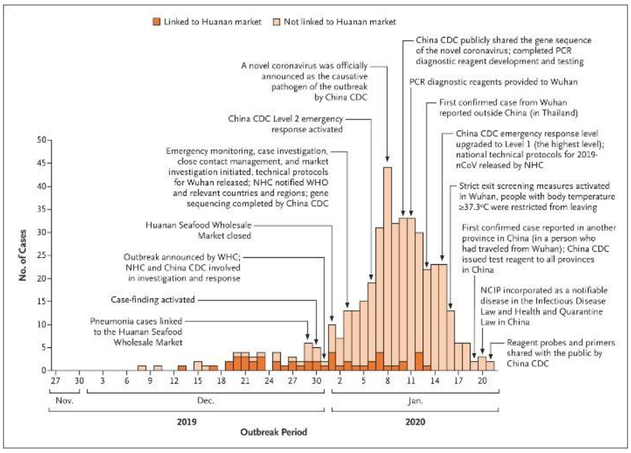


Fig. 1. Incidence trend of the first 425 Confirmed Cases of Novel Coronavirus (2019-nCoV) in Wuhan, China.

Fig. 1. shows the distribution of Linked and non-linked COVID19 cases to Huanan Market for the first 425 patients. For the first 425 patients with the confirmed virus, 56% were male. Most cases (55%) with inception before January 1, 2020, were linked to the Huanan Seafood Wholesale Market, compared to 8.6% of the subsequent cases. The report by Ying Liu et al [10] showed that the reproductive number (R_0) of COVID19 which indicates the transmissibility of a virus is higher when compared to SARS coronavirus.

The limitation to some of these studies on COVID19 is the use of insufficient data and short time to make adequate conclusions. This therefore will make their estimates such as the reproductive number of COVID19 to be possibly biased. However, we believe with more data gathered, it is expected that the estimation error can decrease and provide better visibility of the virus.

III. METHODOLOGY

The Knowledge in Database (KDD) methodology was used for the data analysis, database integration and knowledge gathering of this research.

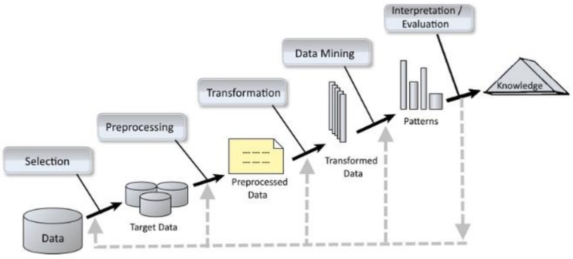


Fig. 2. Methodology for Data Analysis

Fig. 2. shows the steps for the gathering of the data, until knowledge was obtained from the analysed data.

The following three technologies below were selected and applied to successfully to carry out end to end analytics of this research project.

- ***Python Programming Language***

The object-oriented programming language offers a variety of libraries which makes data gathering easy and achievable. Also, analytics tasks were carried out using this language

- ***Mongo DB***

MongoDB is a database system that was used to store the non-relational data. It provides data scalability and is high in performance. MongoDB is document oriented and can store data in the form of collections and it basically generates unique object Id for individual record. It easily handles JSON data and it is schema free.

- ***PostgreSQL***

PostgreSQL was used to manage the relational database system which stored the data in tabular format (structured data). PostgreSQL offers user defined data types, multi-version concurrency control, table inheritance, advance locking system, advance performance optimization etc. It is designed to be extensible for relational queries. It also supports SQL and for Non-relational queries, it supports JSON.

1. DATA SELECTION

The four datasets were obtained from the following source links shown below. Python programming language was used to extract information from the sources and load into MongoDB for storage.

a) XML Dataset

The obtained data in **xml file** format is termed Dataset1. The dataset captures information about the geographic distribution of COVID-19 cases worldwide. This dataset contains 10 columns and 9514 rows in a xml format. The variables are *Date of Reporting*, *Day*, *Month*, *Year*, *Cases*, *Deaths*, *Countries*, *GeoID*, *Country Territory Code and Population*. More details about the attributes of this dataset can be found in the appendix section of this research paper.

- ***Source:***

https://opendata.ecdc.europa.eu/covid19/casedistribution/xml/

b) JSON Dataset

The obtained data in **json file** format is termed Dataset2. The dataset was fetched from twitter by using hashtag of ***#covid19, #lockdown, #socialdistance***. The JSON file was also obtainable from Kaggle. It contains 1000 records. More details about the attributes of this dataset can be found in the appendix section of this research paper.

- ***Source:*** https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

c) Web Scrapping Dataset

The data obtained via **web scrapping** is termed Dataset3. Two sets of data were scrapped here. First piece of data was scrapped off worldometer, a web page with live feed of daily statistics of cases with covid-19 across all countries in the world. This data has 219 records, while the second piece of data was scrapped by using an API-Call to extract the data about hospital information for Medicare in USA. This data has 2000 records. More details about the attributes of this dataset can be found in the appendix section of this research paper.

- ***Source a:*** https://www.worldometers.info/coronavirus/.
- ***Source b:*** https://data.medicare.gov/widgets/xubh-q36u

d) CSV File Dataset

The obtained data in **csv file** format is termed Dataset4. Three sets of data were obtained here. First piece contains detailed information about vi. The dataset contains 13173 records. It captures description of Covid19 patients from January 2020. The second piece contains information about visitors to Wuhan city during the breakout of the virus. It contains record from January to April 2020. Third piece contains a time series information about confirmed, recovered and death cases of the virus. This dataset has 750 records (250 records for each case). More details about the attributes of this dataset can be found in the appendix section of this research paper.

- ***Source:*** https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

2. DATA PRE-PROCESSING

All the data retrieved from MongoDB were pre-processed in a series of steps.

- We identified some null (NA) values, which were removed, and some irrelevant columns were dropped.
- We applied a database technique called Database Normalization for organizing the data in the database. This systematic approach of decomposing tables was used to eliminate redundant (useless) data and ensure data dependencies makes sense i.e. data is logically stored.
- We created a separate table for each set of related data.
- We created unique IDs to identify each set of related data with unique primary keys.
- We also applied log function to make the data for analysis normalized.

3. DATA TRANSFORMATION

Initially, we stored the XML, CSV, JSON, Web scrapping file in MongoDB. We retrieved the files, cleaned the files and fetched the relevant columns in each file. We created five tables with relevant columns.

- Table *COVID19_XML_TBL* was used to store XML file
- Table *COVID19_WEB_TBL* was to store the web scrapping data.
- Table *TWEET_JSON_TBL* was used to store the twitter data.
- Table *COUNTRY_WUHAN_RELATION_CSV_TBL* was used to store the csv file.
- Table *USA_HOSPITAL_WEB_TBL* was used to store the web scrapping data about hospital information.

The cleaned and transformed was saved in PostgreSQL database for storage and quick access.

Database schema diagram, which can also be referred to as entity-relationship diagram shows the relationships among entities/tables stored in the database. The tables created had a "*many to one*" relationship.

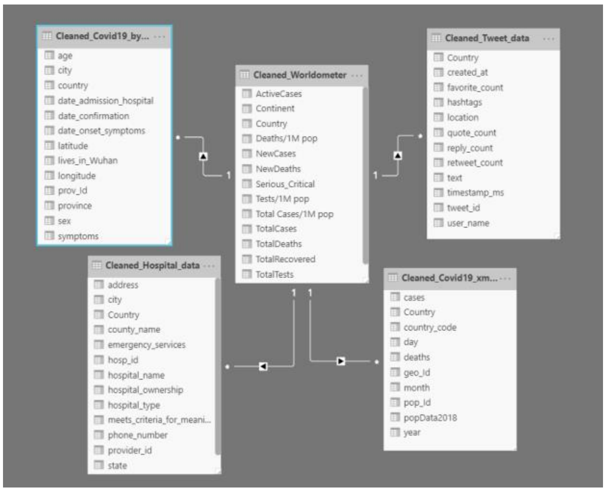


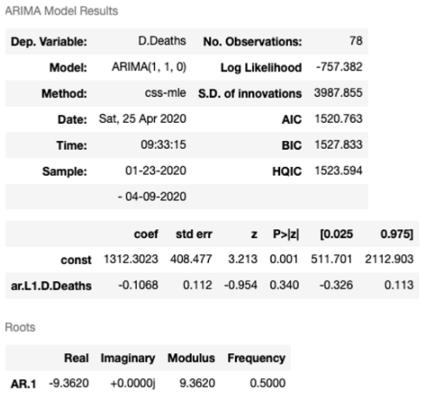
Fig. 3. Database Schema Diagram

Fig. 3. shows the database schema diagram for the created database of the project.

4. DATA MINING

The Autoregressive Integrated Moving Average model (ARIMA) was used to carry out a time series forecast on the data. We used ARIMA model to plot the forecast of prediction of the number of incidence cases, number of recovered cases and number of death cases. The ARIMA model has three key attributes:

p = number of lags / orders of AR terms
d = order of differencing
q = number of lagged forecast errors / order of MA terms.



ARIMA Model Results

Dep. Variable:	D.Deaths	**No. Observations:**	78
Model:	ARIMA(1, 1, 0)	**Log Likelihood**	-757.382
Method:	css-mle	**S.D. of innovations**	3987.855
Date:	Sat, 25 Apr 2020	**AIC**	1520.763
Time:	09:33:15	**BIC**	1527.833
Sample:	01-23-2020	**HQIC**	1523.594
	- 04-09-2020		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1312.3023	408.477	3.213	0.001	511.701	2112.903
ar.L1.D.Deaths	-0.1068	0.112	-0.954	0.340	-0.326	0.113

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-9.3620	+0.0000j	9.3620	0.5000

Fig. 4. ARIMA Model Summary

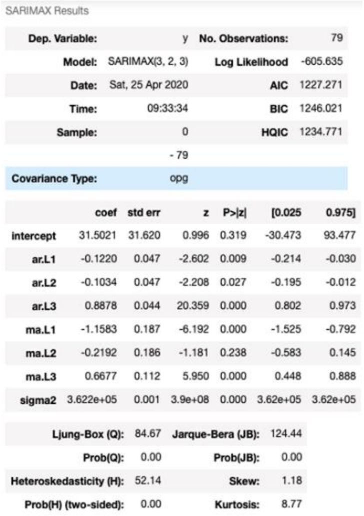
Fig. 4. shows the ARIMA model summary. Here, we can see that AIC is 1520.763. We tried to reduce AIC by using best optimal value of p, d and q.

To successfully make the best predictions, we used Auto-ARIMA model to identify the best optimal value of p, d and q. In ARIMA model, the Akaike Information Criterion (AIC) values signifies how much relative information has been lost by the model. We used a 95% confidence interval for the prediction.

5. DATA INTERPRETATION

The forecast model predicted a significant increase in the death case from coronavirus around the world.

We were able to reduce AIC from 1520.763 to 1227, which is quite a significant improvement from the last model information loss.



SARIMAX Results

Dep. Variable:	y	**No. Observations:**	79
Model:	SARIMAX(3, 2, 3)	**Log Likelihood**	-605.635
Date:	Sat, 25 Apr 2020	**AIC**	1227.271
Time:	09:33:34	**BIC**	1246.021
Sample:	0	**HQIC**	1234.771
	- 79		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	31.5021	31.620	0.996	0.319	-30.473	93.477
ar.L1	-0.1220	0.047	-2.602	0.009	-0.214	-0.030
ar.L2	-0.1034	0.047	-2.208	0.027	-0.195	-0.012
ar.L3	0.8878	0.044	20.359	0.000	0.802	0.973
ma.L1	-1.1583	0.187	-6.192	0.000	-1.525	-0.792
ma.L2	-0.2192	0.186	-1.181	0.238	-0.583	0.145
ma.L3	0.6677	0.112	5.950	0.000	0.448	0.888
sigma2	3.622e+05	0.001	3.9e+08	0.000	3.62e+05	3.62e+05

Ljung-Box (Q):	84.67	**Jarque-Bera (JB):**	124.44
Prob(Q):	0.00	**Prob(JB):**	0.00
Heteroskedasticity (H):	52.14	**Skew:**	1.18
Prob(H) (two-sided):	0.00	**Kurtosis:**	8.77

Fig. 5. Sumary of SARIMAX Results

Fig. 5. shows the summary of the Auto ARIMA model. The model selected the best optimal value of p,d,q i.e. three values in ARIMA model.

IV. RESULTS

1. Visualizations of Data Analysis

Analysis carried out with the data gathered from the different source, provided some knowledge insights. Various visualizations depicting these analytical results are shown below.

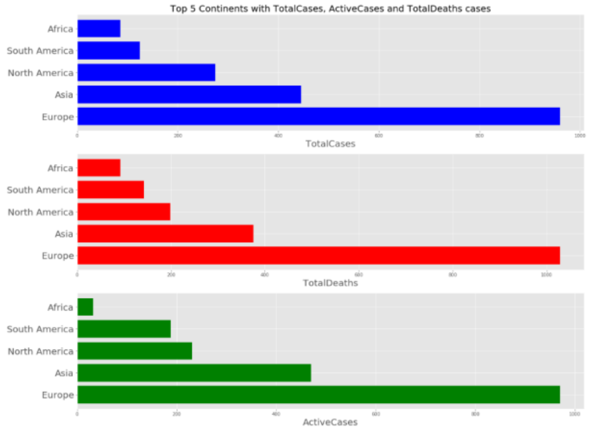


Fig. 6. Top 5 continents with COVID19 cases

Fig. 6. shows the top 5 continents with total cases, active cases and total deaths of COVID19

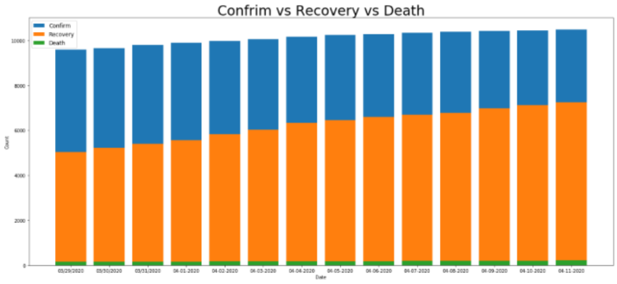


Fig. 7. Bar plot of Confirmed vs Recovered vs Death Cases

Fig. 7. shows the trend for confirmed cases, recovered cases and death cases for the first 14 days in April 2020.

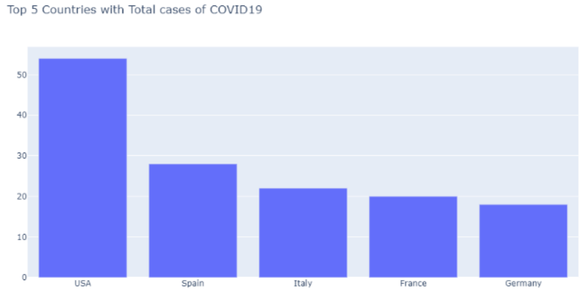


Fig. 8. Top 5 Countries of the world with COVID19 cases

Fig. 8. shows the top 5 countries with total cases of COVID19. USA currently has the highest total cases of COVID19 across all countries of the world. This therefore led to exploration of available hospitals in the country.

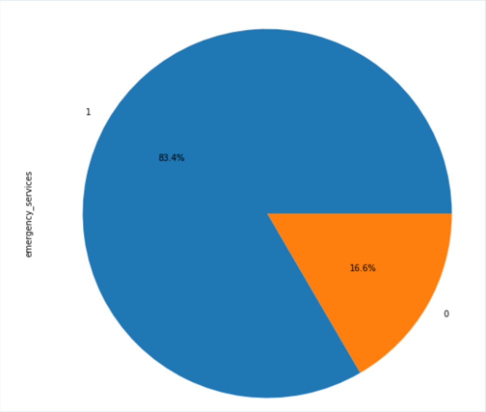


Fig. 9. Percentage of Hospital with emergency services in USA

Fig. 9. shows that 83.4% of hospitals in USA have emergency services. This shows there are large number of hospitals in USA with emergency services that can attend to emergency cases of coronavirus outbreak in the country.

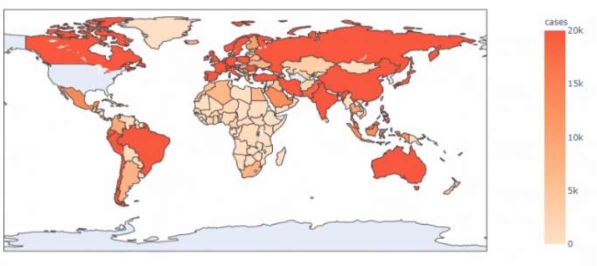


Fig. 10. World map plot of Active Cases in each country

Fig. 10. shows a world map plot that depicts the count of active cases of COVID19 in each country in the form of colour concentration. Currently Europe is the most affected continent.

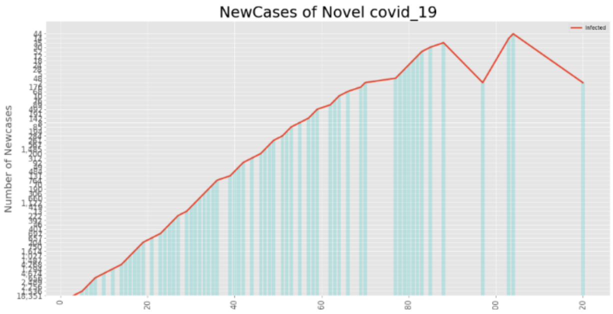


Fig. 11. Plot of increasing trend of coronavirus cases across the world

Fig. 11. shows a line graph that predicts the increase in number of new cases across all countries of the world.

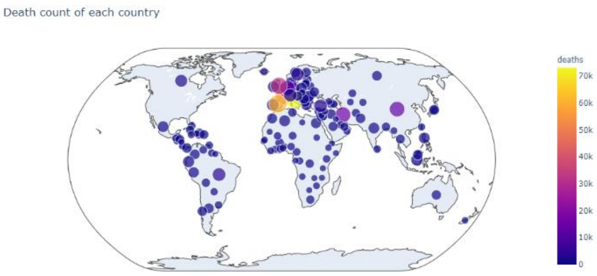


Fig. 12. World map plot of Death count in each country

Fig. 12. shows a bubble graph that depicts the death count of each country and the most affected countries are Italy and France in Europe.

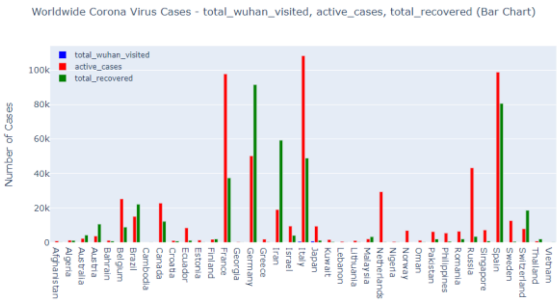


Fig. 13. Active cases and Recovered cases per total Wuhan visited

Fig. 13. shows a bar chart that explains the number of active cases, total recovered cases based on total number of Wuhan visited counts from the specified countries.

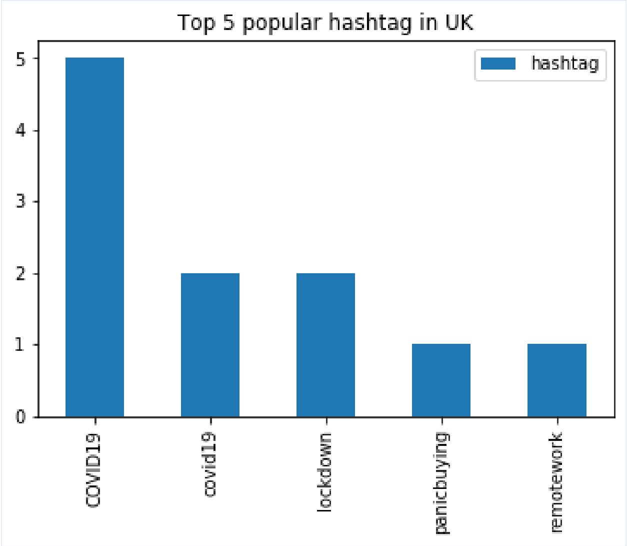


Fig. 14. Popular covid19 Hashtags in UK

Fig. 14. shows that the most popular hashtag on twitter in UK is currently #COVID19. This shows that a lot of people currently follow the conversation about COVID19.

2. Visualizations of Forecasts

Results of the Auto ARIMA model used for forecasts are shown below.

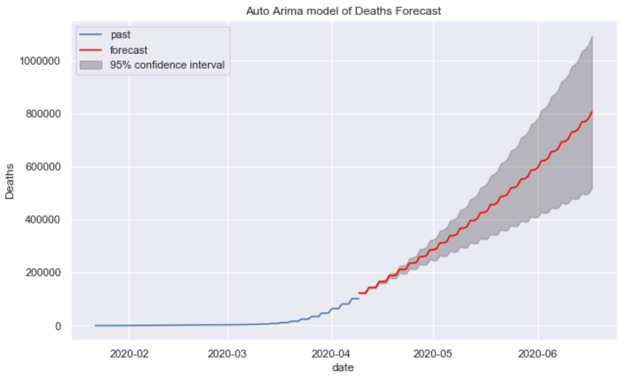


Fig. 14. Forecasts of Death cases using Auto ARIMA

Fig. 14. shows the forecast of death cases using Auto ARIMA model. The number of deaths till mid-March was constant. From April number of deaths has started increasing almost linearly.

The forecasts model predicts that the number of death cases will increase linearly, month after month and very rapidly. This however could be incredible pandemonium across the world with the increase in mortality rate.

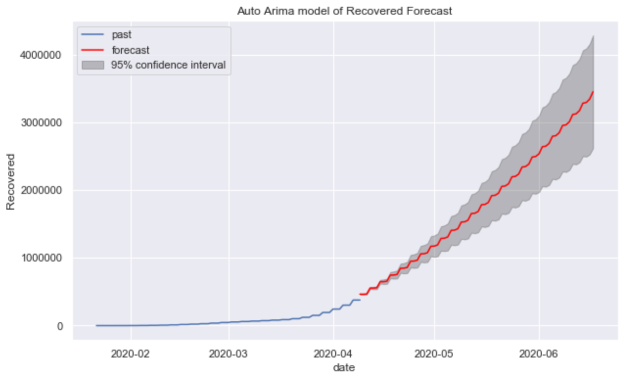


Fig. 15. Forecast of recovered using ARIMA

Fig. 15. shows the forecast of recovered cases using Auto ARIMA. It shows that the number of recovered cases had gradual from February till mid-March, after which a significant linear increase was seen till April.

The forecast model predicts that the number of recovered cases in future will increase linearly, month after month and very rapidly. This is good for humanity and it is due to developing of some good vaccine around the world. This could be more awareness of COVID19, and some good medicine developed by doctors around the world to cure the mild symptoms patients.

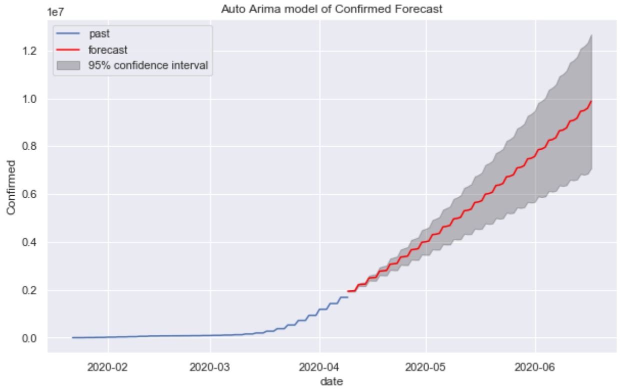


Fig. 16. Forecasts of confirmed cases using ARIMA

Fig. 16. shows the forecast of recovered cases using Auto ARIMA. It shows that the number of confirmed coronavirus cases had gradual increase from February till mid-March, after which a significant linear increase was seen till mid-April.

This forecast predicts that the number of confirmed coronavirus cases will increase linearly month after month. This spread of COVID19 could be multiplied via uncontrolled human to human interaction.

Currently so many people are affected by COVID19, so there is a likelihood of more outbreak of the virus in the future as well.

V. CONCLUSION

USA currently has the highest cases of COVID19. More support and protocols should be implemented to control the spread of the virus.

Analysis carried out on COVID19 epidemic showed that some number of confirmed, recovered and death cases might have been assumed by some governments.

To carry out enough predictive analysis, there will be a need to obtain more attributes of the data, such that more exploratory data analysis can be carried out as well application of adequate models for prediction. To understand the effectiveness as well as the spread of this disease, application of machine learning methods, deep learning and data analytics will play an important role in making models in the future.

It is very important that more awareness is made about COVID19. Also, it is advised that people should take basic hygienic measures and adhere strictly to government protocols all over world in order to contain the spread of the virus. Concepts such as “Social distancing” and use of face mask, should become the new lifestyle to limit spread of the virus.

VI. REFERENCES

[1] Li, Q. et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. N. Engl. J. Med. NEJMoa2001316, https://doi.org/10.1056/NEJMoa2001316 (2020)

[2] ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab (JHU APL) “2019 Novel Coronavirus Visual Dashboard,” 2020. Available from: https://github.com/CSSEGISandData/COVID-19 [Viewed on 24 April 2020]

[3] Jasper Fuk-Woo Chan, MD, Shuofeng Yuan, PhD, Kin-Hang Kok, PhD, Kelvin Kai-Wang To, MD, Hin Chu, PhD, Jin Yang, MD et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster, 2020, VOLUME 395, ISSUE 10223, P514-523, https://doi.org/10.1016/S0140-6736(20)30154-9

[4] Bermingham A, et al. Severe respiratory illness caused by a novel coronavirus, in a patient transferred to the United Kingdom from the Middle East, September 2012. Euro Surveill. 2012; 17(40):20290. https://www.ncbi.nlm.nih.gov/pubmed/23078800

[5] Aleanizy, F.S., Mohmed, N., Alqahtani, F.Y. et al. Outbreak of Middle East respiratory syndrome coronavirus in Saudi Arabia: a retrospective study. BMC Infect Dis 17, 23 (2017). https://doi.org/10.1186/s12879-016-2137-3

[6] Al-Tawfiq JA, Hinedi K, Ghandour J, et al. Middle East respiratory syndrome coronavirus: a case–control study of hospitalized patients. Clinical Infectious Diseases 2014; 59:160–5. https://doi.org/10.1093/cid/ciu226

[7] Durai, P., Batool, M., Shah, M. et al. Middle East respiratory syndrome coronavirus: transmission, virology, and therapeutic targeting to aid in outbreak control. Exp Mol Med 47, e181 (2015). https://doi.org/10.1038/emm.2015.76

[8] Centers for Disease Control and Prevention, Severe Acute Respiratory Syndrome (SARS), 2003, https://www.cdc.gov/sars/.

[9] World Health Organization. Cumulative Number of Reported Probable Cases of Severe Acute Respiratory Syndrome (SARS). Available at: http://www.who.int/csr/sars/country/table2004_04_21/en/.

[10] Ying Liu, Albert A Gayle, Annelies Wilder-Smith, Joacim Rocklöv, The reproductive number of COVID-19 is higher compared to SARS coronavirus, Journal of Travel Medicine, Volume 27, Issue 2, March 2020, taaa021, https://doi.org/10.1093/jtm/taaa021

[11] Data Medicare Gov © 2020. Corona. [Viewed on 9 April 2020]. Available from: https://data.medicare.gov/widgets/xubh-q36u

[12] XML © 2020. [Viewed on 28 March 2020]. Available from: https://catalog.data.gov/dataset/washington-state-public-library-services-affected-by-covid-19

[13] Data.medicare.gov, © 2018. Hospital General Information [Viewed on 12 April 2020]. https://data.medicare.gov/Hospital-Compare/Hospital-General-Information/xubh-q36u

[14] Tarun Kumar, © 2020. COVID-19 Case Study - Analysis, Viz & Comparison [Viewed on 12 April 2020]. https://www.kaggle.com/tarunkr/covid-19-case-study-analysis-viz-comparisons

ABBREVIATIONS

- ***PCR**: Polymerase chain reaction*
- ***WHC**: Wuhan Health Commission*
- ***WHO**: World Health Organization*
- ***NHC**: National Health Commission of the People's Republic of China*
- ***China CDC**: Chinese Centre for Disease Control and Prevention*

APPENDIX

Click icon below

Appendix_Group3_DAP_Project.pdf