Data Science and Machine Learning Integrated Implementation Patterns for Cavernous Knowledge Discovery from COVID-19 Data

B Prashanth¹, G Neelima², Chhaya S Dule³, T Chandra Prakash⁴, S Tarun Reddy⁵

¹,⁴Sumathi Reddy Institute of Technology for Women, Warangal, India.
²Vaagdevi College of Engineering, Warangal, India.
³Dayananda Sagar University, Bangalore, India.
⁵SR Engineering College, Warangal, India.

¹boluk.cse@gmail.com

Abstract: Pandemic is a well-known term for the year 2020. It's essentially a disease that spreads across a region or the entire planet. The entire planet appears powerless, and a jerk triggered by a virus outbreak has halted. On 11 March 2020, WHO announced Corona Virus disease 2019 (COVID-19) a pandemic. The outbreak or epidemic of the virus differs widely from one nation to another. Society is the secret to solving the pandemic. Fever is one of the common, easily detectable symptoms of COVID-19. The COVID 19 in India is one of the most widespread pandemics caused by extreme acute corona viral syndrome 2 (SARS-CoV-2) in coronavirus disease 2019 (COVID-19). On 30 January 2020, the original case of COVID-19 in India, arising in China, was registered. India currently has the highest number of confirmed cases in Asia and the second largest number after the United States of America in the world, with a combined number of confirmed cases exceeding the thresholds of 100,000 on 19 May and 1,000,000 confirmed cases on 17 July 2020. The largest one-day increase in COVID-19 cases of 78,761 cases was observed in Indian countries on 29 August 2020, surpassing the previous record in US cases of 77,368 on 17 July 2020.

Nowadays, data science tasks are not limited to traditional data analysis with limited attributes and records. In current scenarios, the real-time datasets are huge with enormous attributes and such datasets are very complex to evaluate using classical data analysis tools. For example, the datasets of the medical domain integrate several attributes in which the information of symptoms, diagnosis, travel history, health parameters, and many others are evaluated. To deal with such types of datasets assorted database query tools and programming languages are used.

1. Introduction
Covid-19 is an infectious disease triggered by the extreme acute respiratory syndrome transmitted by droplets of aerosol and touch from an infected person [1, 2]. It is a highly infectious disease that rapidly spreads from person to person across four different routes [3, 4]

1. Aerosol droplets- An aerosol is a suspension in the air of fine solid particles or droplets of oil. Such droplets are released when the infected people cough and sneeze or just speak loudly and trigger infection to anyone who comes in touch.
2. Face-Oral-Routes-The SARS-COV-2 virus is known to spill into the feces, particularly in children, from where it can enter the respiratory tract by touching the mouth or nose with soiled hands. This route of transmission is uncommon.

3. Airborne Route- This mode of transmission, like the influenza virus, remains airborne before it infects the person. This trait of COVID is published in a research paper in Nature Medical.

4. Physical Contact- There is a growing body of evidence to suggest asymptomatic transmission occurs in the community.

There is no vaccine developed at this point to stop this pandemic, that’s why it is speeded so fast causing lots of death due to poor immune systems. This is the first known coronavirus which turned out to be pandemic [5]. The virus infects people in different ways, and its development into a disease is based upon the immune system. Few cases show that people have recovered from virus without even requiring any special treatments, turns out to be immune system capability. People diagnosed with medium critical levels are isolated using self-isolations. A coronavirus is a kind of virus that can cause animals and humans ailments. Coronaviruses’ name has been taken from the Latin word 'corona,’ meaning crown because they are protected by what looks like the royal crown of spiked shell shape. The World Health Organization (WHO) has formally announced the discovery of a new virus that will be named by 2019-nCoV in January 2020. The virus has been identified as part of the coronavirus community involving SARS and the other known colds [6].

China was the first country to get infected by the virus, below is the image where the impact of lockdown is quite visible. We can see in the below graph that the number of patients drastically decreased after imposing strict lockdown. Infect it was the discipline of the people and government that despite being the world’s most populous country the virus wasn’t that widespread. There are 82,880 cases reported in china out of which 77,766 are recovered and 4,633 deaths have been reported. Produces a disease that is authoritatively known as COVID-19 and it has spread to at least 141 nations and regions, dying more than 5,700 people worldwide. Someone infected by coronavirus will show common symptoms like fever, tiredness, dry cough, shortness of breath, aches, and pains, sore throat, or diarrhea. However, some people infected with the virus do not show any symptoms and do not feel uncomfortable. Approximately 80% of people infected with COVID-19 may recover without specific treatment, but it is so dangerous for the elderly or someone with serious illness. Which is higher their risk of having a serious illness and experiencing trouble breathing? Right now, no effective COVID 19 vaccine has been created or specific medication has been developed for the treatment of such a virus. Potential vaccines and other unique drug therapies are also under study and are now being extensively tested by leading medical research centers. Besides, the WHO is organizing exceptional efforts to create and manufacture successful vaccines and drugs to prevent and treat COVID-19. As the recorded death case and infected people continue to increase, lock-downs have been undertaken by several nations to reduce the spreading impact of the coronavirus. They often seek to classify the infected among crowds by using an infrared thermometer to measure the temperature in public spaces [7,8,9,10].

But, there are still a lack of use of the infrared thermometer gun itself as it does not reach to all people and time-consuming. This way may also allow the virus to spread widely as the health officer has to do so one by one through a lot of people queueing because one of them is likely to infect people around it. An alternate technology is required to prevent this flaw. The Internet of Things (IoT) has been embraced as the gateway to connectivity in a developing city after it implemented a smart city concept. The main research efforts that are currently being carried out are IoT-based proof of the monitoring of development by remote sensing. The Internet of Things (IoT) is the interconnection between the physical objects or items connected to sensors and software for collecting and transmitting information between them and primary servers with the least human mediation. IoT healthcare is a futuristic concept conveying the facilities and the relevant medical details even more. The medical IoT network is now in an advanced configuration that involves so many variations of systems such as
smart sensors, medical equipment, big data, cloud computing, telemedicine, clinical information system, and much more. IoT techniques are classified as: remote patient tracking, remote health monitoring, and control, sensor-based hand washes monitoring systems, and integrated RFID monitoring [11,12,13].

It provides the best assessment, better diagnosis, and retains patient successful care. IoT's key uses in the area of healthcare include. Administration of medical results, Telemedicine, and portable health care and health management are very important. This study aims at designing a device with the potential to detect the coronavirus automatically from the thermal image with fewer human interactions using a smart mask with Mounted Thermal Imaging System with the IoT technology that has been widely applied in the healthcare sector. The thermal camera technology is built into the smart mask and combined with IoT technology to monitor the screening process to get the data in real-time.

A wide variety of tools and technologies are available which are used by the data analysts and data scientists. In this way, effective prediction and visualization can be done for decision making [14,15].

The following are the commonly used tools for Data Exploration and Analytics for different data analysis tasks

| Tool       | URL                                      |
|------------|------------------------------------------|
| Trifacta   | https://www.trifacta.com/start-wrangling/|
| RapidMiner | https://rapidminer.com/                  |
| Rattle     | https://cran.r-project.org/bin/windows/base/|
| QlikView   | http://global.qlik.com/us/landing/go-sm/qlikview/download-qlikview |
| Weka       | https://www.cs.waikato.ac.nz/ml/weka/    |
| KNIME      | https://www.knime.org/knime-analytics-platform |
| Orange     | http://orange.biolab.si/                |
| Tableu Public | https://public.tableau.com/s/        |
| Open Refine| https://openrefine.org/                  |
| Talend     | https://www.talend.com/products/data-preparation/#free-desktop |
| Data Preparator | http://www.datapreparator.com/downloads.html |
| Tanagra    | http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html |
| H2O        | http://www.h2o.ai/download/h2o/choose   |
| Gephi      | https://gephi.org/                      |

In addition to the abovementioned tools and suites, the programming languages are also used in which statistical and database handling can be done with high performance. These include Python, Julia, R, and many others. It depends on the application area and dataset to decide whether to use high-level language or any tool for data science. The data scientists use the tools or programming languages as per the expectations and outcomes from the data science task.

2. Using Python Pandas for Database Implementations

Python is one of the very powerful, multi-featured, cross-platform, and high-performance programming languages that is used for multiple applications including Data Science, Machine Learning, Deep Learning, Cyber Security, Cloud Computing, Grid Computing, Internet of Things, Parallel Computing, Social Media Mining and many others [16, 17].
For implementations of data science tasks, Python is having different packages using which predictive mining and analytics can be done with high accuracy. Python is enriched with multiple libraries and packages for data science-based evaluations.

**Working with Data Frames in Python Pandas for Recent Database Applications**

In Python, there is a specialized library with the name Pandas that is used for data analysis and easy to use data processing. The Pandas library is very fast in execution and integrates enormous functions for statistical analysis, data evaluations, database management, and many others.

The data scientists and analysts use the Pandas library as it is having so many features with high accuracy.

Following are the key functionalities in Python Pandas

- Effective Data Frame Management for Datasets
- Loading and Processing of Data in Multiple Formats
- Reshaping of Data for Assorted Research Implementations
- Pivot Table Generation and Visualization
- Grouping and Aggregation of Data
- High-Performance Data Wrangling
- Missing Value Imputations and Data Treatments
- Indexing, Slicing, and Subsetting of Large datasets
- Melting, Joining, Concatenation, and Merging
- Time Series Based Analysis and Predictions

**3. Data Analysis on COVID-19 Dataset using Python Pandas**

As in the current situation, the world is facing the menace of COVID-19 and huge loss of human lives is occurring worldwide. The medical experts as well as data scientists are doing restless efforts to analyze the patterns of the COVID-19 virus using different techniques [18, 19].

The datasets of COVID-19 are available on the public domain for the researchers and academicians so that the solutions and predictions can be worked out by different experts [20-24].

Following are a few of the links on the datasets of COVID-19 which are publically available for data scientists and researchers:

**COVID-19 Datasets for Data Science Based Research**

- data.humdata.org/dataset
- ieee-dataport.org/open-access/corona-virus-covid-19-tweets-dataset
- data.world/datasets/covid-19
- kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset
- kaggle.com/imdevskp/corona-virus-report
- github.com/datasets/covid-19
- dimensions.ai/news/dimensions-is-facilitating-access-to-covid-19-research/
- sirm.org/category/senza-categoria/covid-19/
- dev.to/anujgupta/google-s-25-million-ddatasets-a-perfect-gift-for-aspiring-data-scientists-3ekh
- github.com/CSSEGISandData/COVID-19
- kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/kernels
- data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases
- github.com/ieee8023/covid-chestxray-dataset
- github.com/CSSEGISandData/COVID-19
- github.com/ieee8023/covid-chestxray-dataset
For the upcoming data analysis tasks, the dataset in Comma Separated Values (CSV) format of COVID is downloaded and then evaluated using Python Pandas. These datasets can be directly imported in the Data Frames using Python Pandas. The Data Frames in Python Pandas are used to process high-performance operations on the datasets.

### Grouping and Aggregation of Data on COVID Dataset

In the actual dataset shown in Figure 1, there are thousands of records with date wise entries. On each date, the data of patients are recorded.

![Figure 1: Dataset of COVID-19 with Date-Wise Records](image)

First of all, the dataset is imported in the Data Frame which is available in Python Pandas. Following is a very small code of Python Pandas that is executed to perform the grouping of data from the COVID dataset after importing the CSV dataset in Data Frame.

```python
import pandas as pd
dataset = pd.read_csv("covid_19_data.csv")
dval = dataset.values
print (dataset)
grouped=dataset.groupby('State')
print(grouped.get_group('Hebei'))
grouped.get_group('Hebei').to_csv('n.csv')
```

First of all, the dataset is imported in the Data Frame which is available in Python Pandas. Following is a very small code of Python Pandas that is executed to perform the grouping of data from the COVID dataset after importing the CSV dataset in Data Frame.
With the execution of code, the grouping is done by State and then presented in an understandable and grouped format as in Figure 2 above. With this approach, the date-wise-grouping is performed so that the huge dataset can be confined and can be presented effectively.

Melting of Large Dataset using Pandas

Another task that is being performed is related to melting or decomposing the huge dataset. In the dataset shown in Figure 3, the Age, Gender, and Death (Yes / No) is recorded. The dataset from other formats can also be imported in Data Frame of Python Pandas. Once the data is imported to a Data Frame, then different operations can be done. Following is the code snippet of data melting on the COVID dataset.

```python
import pandas as pd
dataset = pd.read_csv("covid.csv")
print (dataset)
d2=pd.melt(dataset, id_vars =['Age'], value_vars =['Gender'])
print (d2)
d3=pd.melt(dataset, id_vars =['Died'], value_vars =['Gender'])
print (d3)
```

With the execution of code, the grouping is done by State and then presented in an understandable and grouped format as in Figure 2 above. With this approach, the date-wise-grouping is performed so that the huge dataset can be confined and can be presented effectively.

Melting of Large Dataset using Pandas

Another task that is being performed is related to melting or decomposing the huge dataset. In the dataset shown in Figure 3, the Age, Gender, and Death (Yes / No) is recorded. The dataset from other formats can also be imported in Data Frame of Python Pandas. Once the data is imported to a Data Frame, then different operations can be done. Following is the code snippet of data melting on the COVID dataset.

```python
import pandas as pd
dataset = pd.read_csv("covid.csv")
print (dataset)
d2=pd.melt(dataset, id_vars =['Age'], value_vars =['Gender'])
print (d2)
d3=pd.melt(dataset, id_vars =['Died'], value_vars =['Gender'])
print (d3)
```

With the execution of code, the grouping is done by State and then presented in an understandable and grouped format as in Figure 2 above. With this approach, the date-wise-grouping is performed so that the huge dataset can be confined and can be presented effectively.

Melting of Large Dataset using Pandas

Another task that is being performed is related to melting or decomposing the huge dataset. In the dataset shown in Figure 3, the Age, Gender, and Death (Yes / No) is recorded. The dataset from other formats can also be imported in Data Frame of Python Pandas. Once the data is imported to a Data Frame, then different operations can be done. Following is the code snippet of data melting on the COVID dataset.

```python
import pandas as pd
dataset = pd.read_csv("covid.csv")
print (dataset)
d2=pd.melt(dataset, id_vars =['Age'], value_vars =['Gender'])
print (d2)
d3=pd.melt(dataset, id_vars =['Died'], value_vars =['Gender'])
print (d3)
```
Figure 3: COVID-19 Dataset with Multiple Attributes

|   |   |   |   |   |
|---|---|---|---|---|
| 1 | Age | Weight | Died | Gender |
| 1 | 19  | 0     |   0  | m     |
| 3 | 35  | 0     |   0  | m     |
| 4 | 26  | 0     |   0  | m     |
| 5 | 17  | 0     |   0  | f     |
| 6 | 19  | 0     |   1  | f     |
| 7 | 23  | 0     |   0  | f     |
| 8 | 27  | 0     |   0  | f     |
| 9 | 32  | 1     |   0  | m     |
| 10| 42  | 0     |   0  | m     |
| 11| 35  | 0     |   0  | f     |
| 12| 26  | 0     |   0  | m     |
| 13| 35  | 0     |   0  | f     |

Figure 4: Melted Dataset of COVID-19 on Association of Variables

As depicted in Figure 4, after melting the dataset can be used for machine learning tasks and data science-based evaluations. In the processes of melting, the mapping of Age and Gender is done separately and the association of Gender with Death is presented. Using this approach, the large dataset can be broken down for further implementations of data science and research tasks with controlled variables.

4. Conclusion
COVID 19 is a new disease that is being studied in several aspects of its dissemination. It spreads across humans faster than influenza, but not as quickly as measles. Individuals are more contagious when they undergo symptoms (even minor or unspecific), but also contagious for up to two days (pre-symptomatic transmission). It is predicted that in mild cases it stays contagious for between seven to 12 days and in extreme cases on average two weeks. The virus can also be spread without any symptoms (asymptomatic transmission), but how much this occurs is not clear. In a June 2020 study, 40-45% of infected persons were asymptomatic. As the datasets of COVID-19, as well as other viruses, are available in the public domain, these can be used by the data scientists and researchers for identification and prediction of root causes of the disease. Such datasets can be trained using machine learning and deep learning models for predictive mining and analysis for contribution to society.
5. References

[1] Alimadadi, A, Aryal, S, Manandhar, I, Munroe, P B, Joe, B and Cheng, X 2020 Artificial intelligence and machine learning to fight COVID-19

[2] Ardabili, S F, Mosavi, A, Ghamisi, P, Ferdinand, F, Varkonyi-Koczy, A R, Reuter, U and Atkinson, P M (2020) Covid-19 outbreak prediction with machine learning Available at SSRN 3580188

[3] Tuli, S, Tuli, S, Tuli, R, & Gill, S S (2020) Predicting the Growth and Trend of COVID-19 Pandemic using Machine Learning and Cloud Computing Internet of Things, 100222

[4] Debnath, S, Barnaby, D P, Coppa, K, Makhnevich, A, Kim, E J, Chatterjee, S, & Hirsch, J S (2020) Machine learning to assist clinical decision-making during the COVID-19 pandemic Bioelectronic medicine 6(1) 1-8

[5] Malki, Z, Atlam, E S, Hassanien, A E, Dagnew, G, Elhosseini, M A, & Gad, I (2020) Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches Chaos Solitons & Fractals 138 110137

[6] Pinter, G, Felde, I, Mosavi, A, Ghamisi, P, & Gloaguen, R (2020) COVID-19 Pandemic Prediction for Hungary; a Hybrid Machine Learning Approach Mathematics 8(6) 890

[7] Elaziz, M A, Hosny, K M, Salah, A, Darwish, M M, Lu, S, & Sahlol, A T (2020) New machine learning method for image-based diagnosis of COVID-19 Plos one 15(6) e0235187

[8] Pasha SN, Harshavardhan A, Ramesh D and Md S Shabana 2019 Variation analysis of artificial intelligence machine learning and advantages of deep architectures International Journal of Advanced Science and Technology 28(17) 488-495

[9] Bikku T Rao NS and Akepogu AR 2019 A novel multi-class ensemble model based on feature selection using Hadoop framework for classifying imbalanced biomedical data International Journal of Business Intelligence and Data Mining 14(43832) 25-39 101504/IJBIDM2019096801

[10] Jayanthi N, Babu BV and Rao NS 2017 Survey on clinical prediction models for diabetes prediction Journal of Big Data 4(1) - 101186/s40537-017-0082-7

[11] Lalmuanauma, S, Hussain, J, & Chhakchhuak, L (2020) Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review Chaos Solitons & Fractals, 110059

[12] Seena Naik K and Sudarshan E 2019 Smart healthcare monitoring system using raspberry Pi on IoT platform ARPN Journal of Engineering and Applied Sciences 14(4) 872-876

[13] Deepak, N, Rajendra Prasad, C & Sanjay Kumar, S 2018, "Patient health monitoring using IoT", International Journal of Innovative Technology and Exploring Engineering, vol 8 no 2 pp 454-457

[14] Ramesh D, Md S Pasha SN, Harshavardhan A and Shabana 2019 Enhancements of artificial intelligence and machine learning International Journal of Advanced Science and Technology 28(17) 16-23

[15] Vaid, S, Cakan, C, & Bhandari, M (2020) Using machine learning to estimate unobserved COVID-19 infections in North America The Journal of bone and joint surgery American volume

[16] Roy, A N, Jose, J, Sunil, A, Gautam, N, Nathalia, D, & Suresh, A (2020) Prediction and Spread Visualization of Covid-19 Pandemic Using Machine Learning

[17] Randhawa, G S, Soltysiak, M P, El Roz, H, de Souza, C P, Hill, K A, & Kari, L (2020) Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study Plos one 15(4) e0232391

[18] Souza, F S H, Hojo-Souza, N S, Santos, E B, Silva, C M, & Guidoni, D L (2020) Predicting the disease outcome in COVID-19 positive patients through Machine Learning: a retrospective cohort study with Brazilian data medRxiv
[19] Nemati, M, Ansary, J, & Nemati, N (2020) Machine-Learning Approaches in COVID-19 Survival Analysis and Discharge-Time Likelihood Prediction Using Clinical Data Patterns 1(5) 100074

[20] Mahender K, Kumar TA and Ramesh KS 2019 Simple transmit diversity techniques for wireless communications Advances in Intelligent Systems and Computing 669 329-342 10.1007/978-981-10-8968-8_28

[21] Rajasri I, Gupta AVSSKS and Rao YVD 2016 Generation of Egts: Hamming Number Approach Procedia Engineering 144 537-542 10.1016/j.proeng.2016.05.039

[22] Wang, P, Zheng, X, Li, J, & Zhu, B (2020) Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics Chaos, Solitons & Fractals 139 110058

[23] Sudarshan E, Naik K.S, Kumar P.P 2020 Parallel approach for backward coding of wavelet trees with CUDA. ARPN Journal of Engineering and Applied Sciences 15(9), pp.1094-1100

[24] Xiang, X, Lu, X, Halavanau, A, Xue, J, Sun, Y, Lai, P H L, & Wu, Z (2020) Modern Senicide in the Face of a Pandemic: An Examination of Public Discourse and Sentiment about Older Adults and COVID-19 Using Machine Learning The Journals of Gerontology: Series B
Data Science and Machine Learning Integrated Implementation Patterns for Cavernous Knowledge Discovery from COVID-19 Data

B Prashanth¹, G Neelima², Chhaya S Dule³, T Chandra Prakash⁴, S Tarun Reddy⁵

¹,⁴Sumathi Reddy Institute of Technology for Women, Warangal, India.
²Vaagdevi College of Engineering, Warangal, India.
³Dayananda Sagar University, Bangalore, India.
⁵SR Engineering College, Warangal, India.

boluk.cse@gmail.com

Abstract: Pandemic is a well-known term for the year 2020. It's essentially a disease that spreads across a region or the entire planet. The entire planet appears powerless, and a jerk triggered by a virus outbreak has halted. On 11 March 2020, WHO announced Corona Virus disease 2019 (COVID-19) a pandemic. The outbreak or epidemic of the virus differs widely from one nation to another. Society is the secret to solving the pandemic. Fever is one of the common, easily detectable symptoms of COVID-19. The COVID 19 in India is one of the most widespread pandemics caused by extreme acute corona viral syndrome2 (SARS-CoV-2) in coronavirus disease 2019 (COVID-19). On 30 January 2020, the original case of COVID-19 in India, arising in China, was registered. India currently has the highest number of confirmed cases in Asia and the second largest number after the United States of America in the world, with a combined number of confirmed cases exceeding the thresholds of 100,000 on 19 May and 1.000,000 confirmed cases on 17 July 2020. The largest one-day increase in COVID-19 cases of 78,761 cases was observed in Indian countries on 29 August 2020, surpassing the previous record in US cases of 77,368 on 17 July 2020.

Nowadays, data science tasks are not limited to traditional data analysis with limited attributes and records. In current scenarios, the real-time datasets are huge with enormous attributes and such datasets are very complex to evaluate using classical data analysis tools. For example, the datasets of the medical domain integrate several attributes in which the information of symptoms, diagnosis, travel history, health parameters, and many others are evaluated. To deal with such types of datasets assorted database query tools and programming languages are used.

1. Introduction
Covid-19 is an infectious disease triggered by the extreme acute respiratory syndrome transmitted by droplets of aerosol and touch from an infected person [1, 2]. It is a highly infectious disease that rapidly spreads from person to person across four different routes [3, 4]

1. Aerosol droplets - An aerosol is a suspension in the air of fine solid particles or droplets of oil. Such droplets are released when the infected people cough and sneeze or just speak loudly and trigger infection to anyone who comes in touch.
2. Face-Oral-Routes-The SARS-COV-2 virus is known to spill into the feces, particularly in children, from where it can enter the respiratory tract by touching the mouth or nose with soiled hands. This route of transmission is uncommon.

3. Airborne Route- This mode of transmission, like the influenza virus, remains airborne before it infects the person. This trait of COVID is published in a research paper in Nature Medical.

4. Physical Contact- There is a growing body of evidence to suggest asymptomatic transmission occurs in the community.

There is no vaccine developed at this point to stop this pandemic, that’s why it is speeded so fast causing lots of death due to poor immune systems. This is the first known coronavirus which turned out to be pandemic [5]. The virus infects people in different ways, and its development into a disease is based upon the immune system. Few cases show that people have recovered from virus without even requiring any special treatments, turns out to be immune system capability. People diagnosed with medium critical levels are isolated using self-isolations. A coronavirus is a kind of virus that can cause animals and humans ailments. Coronaviruses' name has been taken from the Latin word 'corona,' meaning crown because they are protected by what looks like the royal crown of spiked shell shape.

The World Health Organization (WHO) has formally announced the discovery of a new virus that will be named by 2019-nCoV in January 2020. The virus has been identified as part of the coronavirus community involving SARS and the other known colds [6].

China was the first country to get infected by the virus, below is the image where the impact of lockdown is quite visible. We can see in the below graph that the number of patients drastically decreased after imposing strict lockdown. Infect it was the discipline of the people and government that despite being the world’s most populous country the virus wasn’t that widespread. There are 82,880 cases reported in china out of which 77,766 are recovered and 4,633 deaths have been reported. Produces a disease that is authoritatively known as COVID-19 and it has spread to at least 141 nations and regions, dying more than 5,700 people worldwide. Someone infected by coronavirus will show common symptoms like fever, tiredness, dry cough, shortness of breath, aches, and pains, sore throat, or diarrhea. However, some people infected with the virus do not show any symptoms and do not feel uncomfortable. Approximately 80 % of people infected with COVID-19 may recover without specific treatment, but it is so dangerous for the elderly or someone with serious illness. Which is higher their risk of having a serious illness and experiencing trouble breathing? Right now, no effective COVID 19 vaccine has been created or specific medication has been developed for the treatment of such a virus. Potential vaccines and other unique drug therapies are also under study and are now being extensively tested by leading medical research centers. Besides, the WHO is organizing exceptional efforts to create and manufacture successful vaccines and drugs to prevent and treat COVID-19. As the recorded death case and infected people continue to increase, lock-downs have been undertaken by several nations to reduce the spreading impact of the coronavirus. They often seek to classify the infected among crowds by using an infrared thermometer to measure the temperature in public spaces [7,8,9,10].

But, there are still a lack of use of the infrared thermometer gun itself as it does not reach to all people and time-consuming. This way may also allow the virus to spread widely as the health officer has to do so one by one through a lot of people queuing because one of them is likely to infect people around it. An alternate technology is required to prevent this flaw. The Internet of Things (IoT) has been embraced as the gateway to connectivity in a developing city after it implemented a smart city concept. The main research efforts that are currently being carried out are IoT-based proof of the monitoring of development by remote sensing. The Internet of Things (IoT) is the interconnection between the physical objects or items connected to sensors and software for collecting and transmitting information between them and primary servers with the least human mediation. IoT healthcare is a futuristic concept conveying the facilities and the relevant medical details even more. The medical IoT network is now in an advanced configuration that involves so many variations of systems such as
smart sensors, medical equipment, big data, cloud computing, telemedicine, clinical information system, and much more. IoT techniques are classified as: remote patient tracking, remote health monitoring, and control, sensor-based hand washes monitoring systems, and integrated RFID monitoring [11,12,13].

It provides the best assessment, better diagnosis, and retains patient successful care. IoT’s key uses in the area of healthcare include. Administration of medical results, Telemedicine, and portable health care and health management are very important. This study aims at designing a device with the potential to detect the coronavirus automatically from the thermal image with fewer human interactions using a smart mask with Mounted Thermal Imaging System with the IoT technology that has been widely applied in the healthcare sector. The thermal camera technology is built into the smart mask and combined with IoT technology to monitor the screening process to get the data in real-time.

A wide variety of tools and technologies are available which are used by the data analysts and data scientists. In this way, effective prediction and visualization can be done for decision making [14,15].

The following are the commonly used tools for Data Exploration and Analytics for different data analysis tasks

| Tool        | URL                                      |
|-------------|------------------------------------------|
| Trifacta    | https://www.trifacta.com/start-wrangling/|
| RapidMiner  | https://rapidminer.com/                  |
| Rattle      | https://cran.r-project.org/bin/windows/base/|
| QlikView    | http://global.qlik.com/us/landing/go-sm/qlikview/download-qlikview|
| Weka        | https://www.cs.waikato.ac.nz/ml/weka/    |
| KNIME       | https://www.knime.org/knime-analytics-platform|
| Orange      | http://orange.biolab.si/                |
| Tableau Public | https://public.tableau.com/s/            |
| Open Refine | https://openrefine.org/                 |
| Talend      | https://www.talend.com/products/data-preparation/#free-desktop |
| Data Preparator | http://www.datapreparator.com/downloads.html |
| Tanagra     | http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html |
| H2O         | http://www.h2o.ai/download/h2o/choose |
| Gephi       | https://gephi.org/                      |

In addition to the abovementioned tools and suites, the programming languages are also used in which statistical and database handling can be done with high performance. These include Python, Julia, R, and many others. It depends on the application area and dataset to decide whether to use high-level language or any tool for data science. The data scientists use the tools or programming languages as per the expectations and outcomes from the data science task.

2. Using Python Pandas for Database Implementations

Python is one of the very powerful, multi-featured, cross-platform, and high-performance programming languages that is used for multiple applications including Data Science, Machine Learning, Deep Learning, Cyber Security, Cloud Computing, Grid Computing, Internet of Things, Parallel Computing, Social Media Mining and many others [16, 17].
For implementations of data science tasks, Python is having different packages using which predictive mining and analytics can be done with high accuracy. Python is enriched with multiple libraries and packages for data science-based evaluations.

Working with Data Frames in Python Pandas for Recent Database Applications
In Python, there is a specialized library with the name Pandas that is used for data analysis and easy to use data processing. The Pandas library is very fast in execution and integrates enormous functions for statistical analysis, data evaluations, database management, and many others.

The data scientists and analysts use the Pandas library as it is having so many features with high accuracy.

Following are the key functionalities in Python Pandas
- Effective Data Frame Management for Datasets
- Loading and Processing of Data in Multiple Formats
- Reshaping of Data for Assorted Research Implementations
- Pivot Table Generation and Visualization
- Grouping and Aggregation of Data
- High-Performance Data Wrangling
- Missing Value Imputations and Data Treatments
- Indexing, Slicing, and Subsetting of Large datasets
- Melting, Joining, Concatenation, and Merging
- Time Series Based Analysis and Predictions

3. Data Analysis on COVID-19 Dataset using Python Pandas
As in the current situation, the world is facing the menace of COVID-19 and huge loss of human lives is occurring worldwide. The medical experts as well as data scientists are doing restless efforts to analyze the patterns of the COVID-19 virus using different techniques [18, 19].

The datasets of COVID-19 are available on the public domain for the researchers and academicians so that the solutions and predictions can be worked out by different experts [20-24].

Following are a few of the links on the datasets of COVID-19 which are publically available for data scientists and researchers:

**COVID-19 Datasets for Data Science Based Research**
- data.humdata.org/dataset
- ieee-dataport.org/open-access/corona-virus-covid-19-tweets-dataset
- data.world/datasets/covid-19
- kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset
- kaggle.com/imdevskp/corona-virus-report
- github.com/datasets/covid-19
- dimensions.ai/news/dimensions-is-facilitating-access-to-covid-19-research/
- sirm.org/category/senza-categoria/covid-19/
- dev.to/anujgupta/google-s-25-million-ddatasets-a-perfect-gift-for-aspiring-data-scientists-3ekh
- github.com/CSSEGISandData/COVID-19
- kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/kernels
- data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases
- github.com/iee8023/covid-chestxray-dataset
- github.com/CSSEGISandData/COVID-19
- github.com/iee8023/covid-chestxray-dataset
For the upcoming data analysis tasks, the dataset in Comma Separated Values (CSV) format of COVID is downloaded and then evaluated using Python Pandas. These datasets can be directly imported in the Data Frames using Python Pandas. The Data Frames in Python Pandas are used to process high-performance operations on the datasets.

Grouping and Aggregation of Data on COVID Dataset

In the actual dataset shown in Figure 1, there are thousands of records with date wise entries. On each date, the data of patients are recorded.

First of all, the dataset is imported in the Data Frame which is available in Python Pandas. Following is a very small code of Python Pandas that is executed to perform the grouping of data from the COVID dataset after importing the CSV dataset in Data Frame.

```python
import pandas as pd
dataset = pd.read_csv("covid_19_data.csv")
dval = dataset.values
print (dataset)
grouped=dataset.groupby('State')
print(grouped.get_group('Hebei'))
grouped.get_group('Hebei').to_csv('n.csv')
```

Figure 1: Dataset of COVID-19 with Date-Wise Records

First of all, the dataset is imported in the Data Frame which is available in Python Pandas. Following is a very small code of Python Pandas that is executed to perform the grouping of data from the COVID dataset after importing the CSV dataset in Data Frame.
With the execution of code, the grouping is done by state and then presented in an understandable and grouped format as in Figure 2 above. With this approach, the date-wise-grouping is performed so that the huge dataset can be confined and can be presented effectively.

**Melting of Large Dataset using Pandas**

Another task that is being performed is related to melting or decomposing the huge dataset. In the dataset shown in Figure 3, the Age, Gender, and Death (Yes / No) is recorded. The dataset from other formats can also be imported in Data Frame of Python Pandas. Once the data is imported to a Data Frame, then different operations can be done. Following is the code snippet of data melting on the COVID dataset.

```python
import pandas as pd
dataset = pd.read_csv("covid.csv")
print(dataset)
d2=pd.melt(dataset, id_vars =['Age'], value_vars =['Gender'])
print (d2)
d3=pd.melt(dataset, id_vars =['Died'], value_vars =['Gender'])
print (d3)
```

Figure 2: Grouping Dataset of COVID-19 on State
As depicted in Figure 4, after melting the dataset can be used for machine learning tasks and data science-based evaluations. In the processes of melting, the mapping of Age and Gender is done separately and the association of Gender with Death is presented. Using this approach, the large dataset can be broken down for further implementations of data science and research tasks with controlled variables.

4. Conclusion
COVID 19 is a new disease that is being studied in several aspects of its dissemination. It spreads across humans faster than influenza, but not as quickly as measles. Individuals are more contagious when they undergo symptoms (even minor or unspecific), but also contagious for up to two days (presymptomatic transmission). It is predicted that in mild cases it stays contagious for between seven to 12 days and in extreme cases on average two weeks. The virus can also be spread without any symptoms (asymptomatic transmission), but how much this occurs is not clear. In a June 2020 study, 40-45% of infected persons were asymptomatic. As the datasets of COVID-19, as well as other viruses, are available in the public domain, these can be used by the data scientists and researchers for identification and prediction of root causes of the disease. Such datasets can be trained using machine learning and deep learning models for predictive mining and analysis for contribution to society.
5. References

[1] Alimadadi, A, Aryal, S, Manandhar, I, Munroe, P B, Joe, B and Cheng, X 2020 Artificial intelligence and machine learning to fight COVID-19

[2] Ardabili, S F, Mosavi, A, Ghamisi, P, Ferdinand, F, Varkonyi-Koczy, A R, Reuter, U and Atkinson, P M (2020) Covid-19 outbreak prediction with machine learning Available at SSRN 3580188

[3] Tuli, S, Tuli, S, Tuli, R, & Gill, S S (2020) Predicting the Growth and Trend of COVID-19 Pandemic using Machine Learning and Cloud Computing Internet of Things, 100222

[4] Debnath, S, Barnaby, D P, Coppa, K, Makhnevich, A, Kim, E J, Chatterjee, S, & Hirsch, J S (2020) Machine learning to assist clinical decision-making during the COVID-19 pandemic Bioelectronic medicine 6(1) 1-8

[5] Malki, Z, Atlam, E S, Hassanien, A E, Dagnew, G, Elhosseini, M A, & Gad, I (2020) Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches Chaos Solitons & Fractals 138 110137

[6] Pinter, G, Felde, I, Mosavi, A, Ghamisi, P, & Gloaguen, R (2020) COVID-19 Pandemic Prediction for Hungary; a Hybrid Machine Learning Approach Mathematics 8(6) 890

[7] Elaziz, M A, Hosny, K M, Salah, A, Darwish, M M, Lu, S, & Sahlol, A T (2020) New machine learning method for image-based diagnosis of COVID-19 Plos one 15(6) e0235187

[8] Pasha SN, Harshavardhan A, Ramesh D and Md S Shabana 2019 Variation analysis of artificial intelligence machine learning and advantages of deep architectures International Journal of Advanced Science and Technology 28(17) 488-495

[9] Bikku T Rao NS and Akepogu AR 2019 A novel multi-class ensemble model based on feature selection using Hadoop framework for classifying imbalanced biomedical data International Journal of Business Intelligence and Data Mining 14(43832) 25-39 101504/IJBIDM2019096801

[10] Jayanthi N, Babu BV and Rao NS 2017 Survey on clinical prediction models for diabetes prediction Journal of Big Data 4(1) - 101186/s40537-017-0082-7

[11] Lalmuanawma, S, Hussain, J, & Chhakchhuak, L (2020) Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review Chaos Solitons & Fractals, 110059

[12] Seena Naik K and Sudarshan E 2019 Smart healthcare monitoring system using raspberry Pi on IoT platform ARPN Journal of Engineering and Applied Sciences 14(4) 872-876

[13] Deepak, N, Rajendra Prasad, C & Sanjay Kumar, S 2018, "Patient health monitoring using IoT", International Journal of Innovative Technology and Exploring Engineering, vol 8 no 2 pp 454-457

[14] Ramesh D, Md S Pasha SN, Harshavardhan A and Shabana 2019 Enhancements of artificial intelligence and machine learning International Journal of Advanced Science and Technology 28(17) 16-23

[15] Vaid, S, Cakan, C, & Bhandari, M (2020) Using machine learning to estimate unobserved COVID-19 infections in North America The Journal of bone and joint surgery American volume

[16] Roy, A N, Jose, J, Sunil, A, Gautam, N, Nathalia, D, & Suresh, A (2020) Prediction and Spread Visualization of Covid-19 Pandemic Using Machine Learning

[17] Randhawa, G S, Soltysiak, M P, El Roz, H, de Souza, C P, Hill, K A, & Kari, L (2020) Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study Plos one 15(4) e0232391

[18] Souza, F S H, Hojo-Souza, N S, Santos, E B, Silva, C M, & Guidoni, D L (2020) Predicting the disease outcome in COVID-19 positive patients through Machine Learning: a retrospective cohort study with Brazilian data medRxiv
[19] Nemati, M, Ansary, J, & Nemati, N (2020) Machine-Learning Approaches in COVID-19 Survival Analysis and Discharge-Time Likelihood Prediction Using Clinical Data Patterns 1(5) 100074

[20] Mahender K, Kumar TA and Ramesh KS 2019 Simple transmit diversity techniques for wireless communications Advances in Intelligent Systems and Computing 669 329-342 10.1007/978-981-10-8968-8_28

[21] Rajasri I, Gupta AVSSKS and Rao YVD 2016 Generation of Egts: Hamming Number Approach Procedia Engineering 144 537-542 10.1016/j.proeng.2016.05.039

[22] Wang, P, Zheng, X, Li, J, & Zhu, B (2020) Prediction of epidemic trends in COVID-19 with logistic model and machine learning techniques Chaos, Solitons & Fractals 139 110058

[23] Sudarshan E, Naik K.S, Kumar P.P 2020 Parallel approach for backward coding of wavelet trees with CUDA. ARPN Journal of Engineering and Applied Sciences 15(9), pp.1094-1100

[24] Xiang, X, Lu, X, Halavanau, A, Xue, J, Sun, Y, Lai, P H L, & Wu, Z (2020) Modern Senicide in the Face of a Pandemic: An Examination of Public Discourse and Sentiment about Older Adults and COVID-19 Using Machine Learning The Journals of Gerontology: Series B