# COVID-19 CASES ANALYSIS USING COGNOS

## ABSTRACT :

Covid-19 is an infectious illness caused by a newly identified form of coronavirus. This is a new virus and illness that was previously unknown before the December 2019 outbreak in Wuhan, China.Therefore this study will discuss the grouping of Cases and Deaths of COVID-19 in EU/EEA Countries. The objective is to compare and contrast the mean values and standard deviations of cases and associated deaths per day and by country in the EU/EEA. This project encompasses defining analysis objectives, collecting COVID-19 data, designing relevant visualizations in IBM Cognos, and deriving insights from the data.The method used is the K-Means Clustering Data Mining and standard deviation using Two path algorithm . By using this method the data that has been obtained can be grouped into several clusters, where K-Means Clustering Process is applied and by following the mean to the standard deviation with the metioned formula.In this phase we will continue building the analysis by creating visualization using IBM cognos and deriving insights from the data.We will create charts and graphs in IBM cognos to visualize and compare the mean values and standard deviation of covid 19 cases. Then we will analyze the visualizations to identify trends, variations and potential correlation between cases and deaths.

## DATASET LINK :

https://www.kaggle.com/datasets/chakradharmattapalli/covid□19-cases

## PREPROCESSING :

Preprocessing of data is a crucial step in data analysis and machine learning. Here are the common steps involved:

### 1.DATA CLEANING:

This involves handling missing values, dealing with duplicates, and correcting errors.

```
miss_mean_imputer = Imputer(missing_values=’NaN’, strategy="mean", axis=0)

miss_mean_imputer = miss_mean_imputer.fit(df)

Imputed_df = miss_mean_imputer.transform(df.values

)
```

```
print(imputed_df)
```

## 2.DATA INTEGRATION:

If data comes from different sources, integration ensures consistency in format and resolving any inconsistencies.

```
df.drop([dateRep],axis=0)
```

## 3.DATA TRANSFORMATION:

This step includes normalization, variables to make data suitable for analysis.

## 4.DATA REDUCTION:

Reducing the dimensionality of the data through techniques like feature selection or efficiency of the analysis.

## 5.DATA DISCRETIZATION:

It continuous variables can be transformed into discrete ones for ease of analysis.

## COMPARISON OF MEAN AND STRANDARD DEVIATION :

### COMPARISON OF STANDARD DEVIATION

The cases having a higher standard deviation tells that the cases is more spread out or dispersed than the deaths. The cases had a standard deviation of 6490.51 whereas the deaths had a standard deviation of 113.9566.

### COMPARISON OF MEAN

The average of cases were greater than average of deaths. The cases having an average of 3661.011 data and the deaths having an average of 65.29194 data.

| | MEAN | STANDARD DEVIATION (SD) |
|---|---|---|
| CASES | 3661.011 | 6490.51 |
| DEATHS | 65.29194 | 113.9566 |

### IN CASES:

# COVID-19 CASES ANALYSIS USING COGNOS

● FRANCE has the highest average of 22206.68 and LIECHTENSTEIN has the lowest average of 4.802197.

○ FRANCE has the highest standard deviation of 13071.98 and

LIECHTENSTEIN has the lowest sd of 4.53681.

**IN DEATH:**

○ POLAND has the highest mean of 329.3297 and ICELAND has the very low mean of 1.0989

● ITALY has the highest standard deviation of 7041.66 and ICELAND has the lowest sd of 1.0483

## METHODOLOGY :

### 1.MEAN USING K-MEANS ALGORITHM :

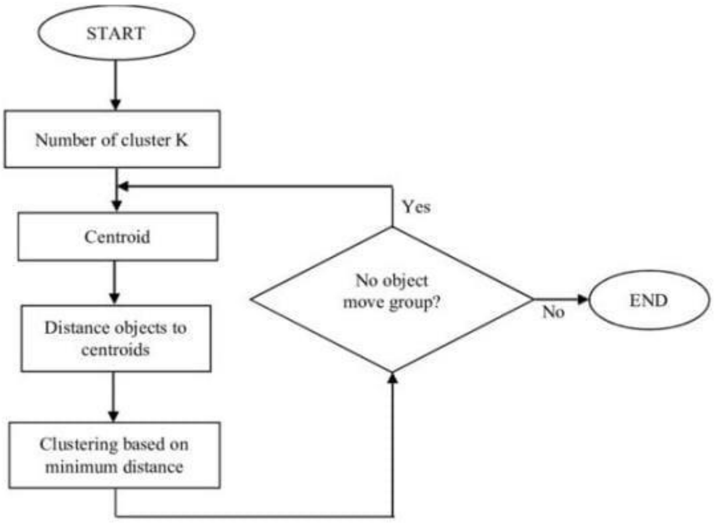
**. RESEARCH METHOD :**

This study uses the K-Means Clustering method. K-Means is one of the clustering algorithms used in the Unsupervised learning group that is used to classify data into several classes with a partition of the system. This algorithm accepts data entries in the form of class labels .

**CENTROID**

A midpoint value, or centroid data, is generated when implementing the K-means algorithm. The method of determining a midpoint value is achieved by following the largest (maximum) for high cluster value (C1), the mean value for a medium cluster (C2), and the lowest Cluster value (C3).

# COVID-19 CASES ANALYSIS USING COGNOS


START
Number of cluster K
Centroid
Distance objects to centroids
Clustering based on minimum distance
No object move group?
Yes
No
END

**RESEARCH FLOWCHART FOR K-MEANS**

## ALGORITHM :

Steps to perform clustering using k-means algorithm :

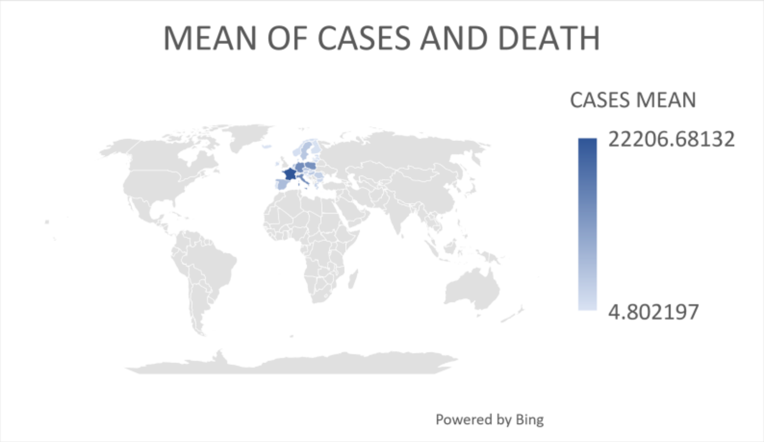
a. Determine cluster counts (k) in the data set.

b. Determine the center value (Centroid).

c. On each record, calculate the closest distance to Centroid

d. Distance Group objects to nearest Centroid

e. Repeat step a to step b, iterating until Centroid is optimal



Let us consider the data segment of country Austria from the dataset, here we taken account of death

[5,6,11,4,19,8,3,3,8,12.........................................................21,23,9,16,24,19,14,17,25

Now we select K random points from the data as centroids

So, No.of clusters into 3 sets named as K1, K2 &K3

K1={3,4,5,6,....13,14,15}

K2={16,17,18,19,....26,27,28}

K3={30,31,32,......,48,46,51}

Assigning all the points to the closest cluster centroids as

M1=(3+4+5+6+......13+14+15)/100

= 9

M2=(16+17+19......+26+27+28)/13

=22

M3=(30+31+32+.....+48+46+51)/19

=37

Compare the first value of k2 with the m values of k1,k2 and k3 then find the difference these three values

If the mean value is same with above step stop orelse repeat the steps

**CONCLUSION FOR MEAN :**

Clustering for Cases and Deaths caused by COVID-19 EU/EEA Countries. clustering uses 3 clusters, that is: (C1) high, (C2) regular cluster, and (C3) low cluster. From the results of clustering, 30 countries is in (C1) High Cluster that is included in the red zone category. The average of cases were greater than average of deaths. The cases having an average of 3661.011 data and the deaths having an average of 65.29194 data. The country France has a highest mean of cases is 22206.68132. The country Liechtenstein has a lowest mean of cases is 4.802197. The country Poland has a highest mean of death is 329.3296703. The country Iceland has a lowest mean of death is 1.09889.

# COVID-19 CASES ANALYSIS USING COGNOS

## 2.STANDARD DEVIATION FROM VARIANCE

Standard deviation measures how far apart numbers are in a data set. Variance, on the other hand, gives an actual value to how much the numbers in a data set vary from the mean. Standard deviation is the square root of the variance and is expressed in the same units as the data set. Thus we use two pass algorithm to find the variance of the data .

## 2.1. VARIANCE BY TWO PASS ALGORITHM :

A two-pass algorithm is a computational method that processes data or performs a task in two sequential passes or steps. Each pass serves a specific purpose and often involves reading, analyzing, or transforming data. Two-pass algorithms are commonly used in various fields of computer science and data processing.

Step 1: finding a mean by mentioned formula. (death in Sweden )

$$\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n},$$

Where , x- death data of sweden country.

n- time period , 90.

Mean=15.96703

Step 2: finding variance by Two pass algorithm.

$$\text{sample variance} = s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Where,

s^2= ((1-15.96703)^2 + (5-15.96703)^2 + (815.96703)^2 +

(215.96703)^2 +........... + (19-15.96703)^2) )/ (90-1)

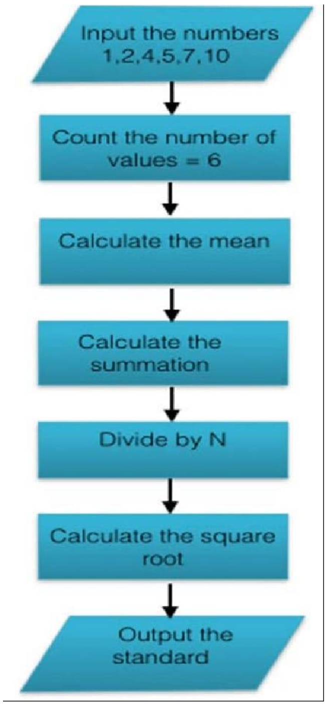
variance = 33.21001

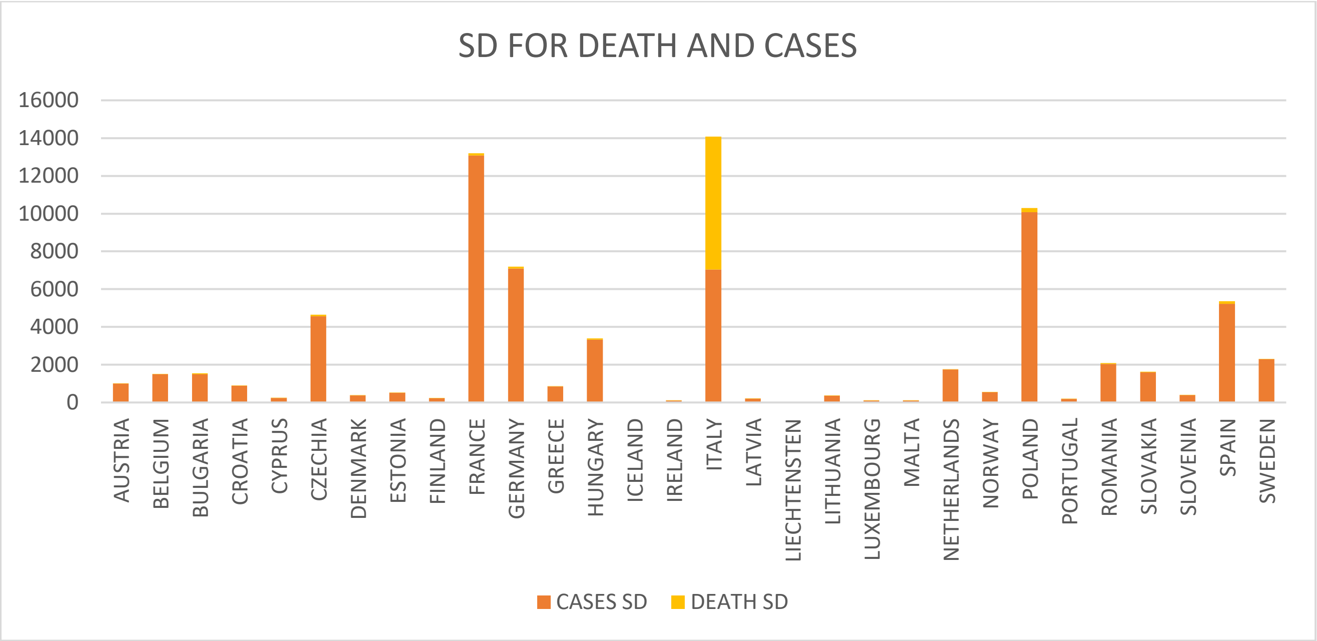
step 3 : standard deviation.

By taking square root for variance we get standard deviation .

SD = 5.762813

**FLOW CHART FOR TWO PASS ALGORITHM**

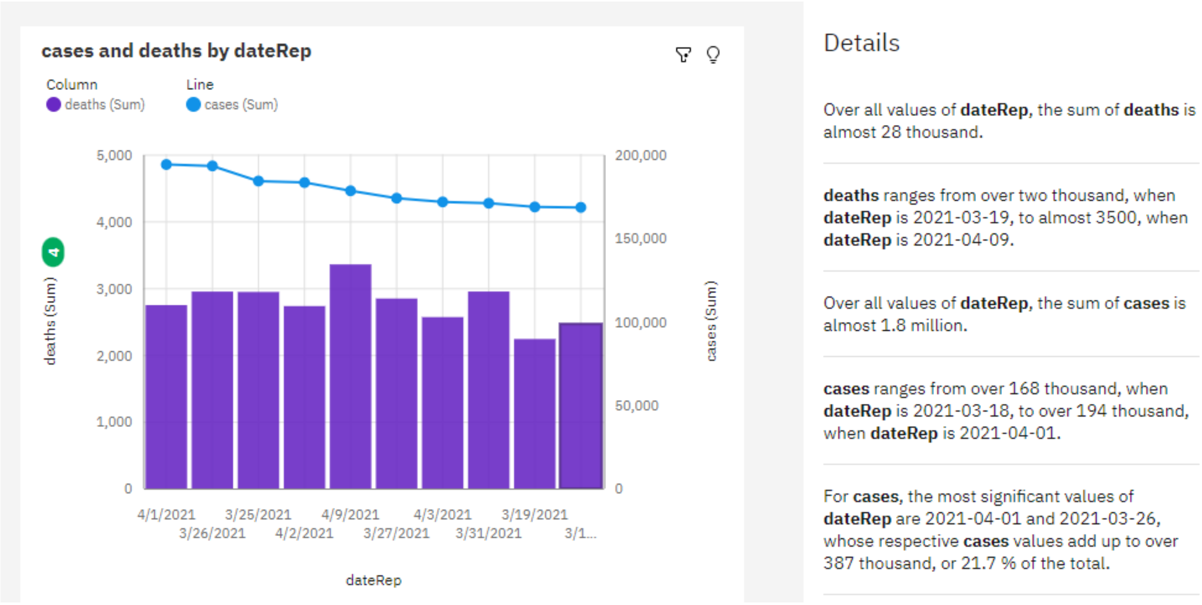

Input the numbers 1,2,4,5,7,10
Count the number of values = 6
Calculate the mean
Calculate the summation
Divide by N
Calculate the square root
Output the standard



**CONCLUSION FOR SD :**

Standard deviation for Cases and Deaths caused by COVID-19 EU/EEA Countries. The cases having an standard deviation of 6490.51 data and the deaths having an standard deviation of 113.9566 data. The country Poland has a highest SD of cases is 10077.11733. The country Iceland has a lowest SD of cases is 8.0349. The country Italy has a highest SD of death is 7041.66. The country Iceland has a lowest SD of death is 1.20483.

## VISUALIZATIONS USING IBM COGNOS :



Details

Over all values of **dateRep**, the sum of **deaths** is almost 28 thousand.

**deaths** ranges from over two thousand, when **dateRep** is 2021-03-19, to almost 3500, when **dateRep** is 2021-04-09.

Over all values of **dateRep**, the sum of **cases** is almost 1.8 million.

**cases** ranges from over 168 thousand, when **dateRep** is 2021-03-18, to over 194 thousand, when **dateRep** is 2021-04-01.

For **cases**, the most significant values of **dateRep** are 2021-04-01 and 2021-03-26, whose respective **cases** values add up to over 387 thousand, or 21.7 % of the total.

## INSIGHTS :

Insights from COVID-19 cases analysis can provide a deeper understanding of the pandemic's impact.

### 1.Diagnostic Methods:

Analyzing the accuracy and availability of testing helps in efficient and widespread diagnosis and contact tracing.

### 2. Mutations and Variants:

Monitoring the emergence of new variants and their potential impact on transmission, vaccine effectiveness, and treatment is essential.

### 3. Epidemiological Patterns:

Understanding how the virus spreads, including the rate of transmission and its seasonality, is crucial for implementing effective control measures.

## TRENDS OF COVID-19 BETWEEN CASES AND DEATHS :

# COVID-19 CASES ANALYSIS USING COGNOS

Trends in COVID-19 cases and deaths have evolved over time, influenced by various factors, including vaccination campaigns, public health measures, new variants, and population behavior. Here are some general trends:

A. **Early Surge in Cases and Deaths:**

At the beginning of the pandemic, there was a rapid increase in both cases and deaths as the virus spread globally without effective countermeasures.

B. **Fluctuations in Cases:**

Throughout the pandemic, there have been waves or surges in cases, often corresponding to changes in public health measures, holiday seasons, and the emergence of new variants.

C. **Lag in Death Trends:**

Deaths tend to lag behind reported cases by several weeks. This is due to the time it takes for individuals to develop severe symptoms and succumb to the virus.

## REGIONAL VARIATIONS:

Analyzing data by regions or countries can reveal variations in the impact of the pandemic. Some areas may have been more heavily affected than others.

## VACCINATION IMPACT:

You can analyze data before and after vaccination campaigns to see how they impact the number of cases and deaths. CASE-FATALITY RATE (CFR): Calculating the CFR, which is the ratio of deaths to cases, can help assess the severity of the virus in a given area or population.

## ACCURACY :

The Jupyter Notebook is an open-source web application that allows you to create and share documents that live code. In Jupyter we use python programming language to find the accuracy of the data with the help of toal number of samples in the dataset.

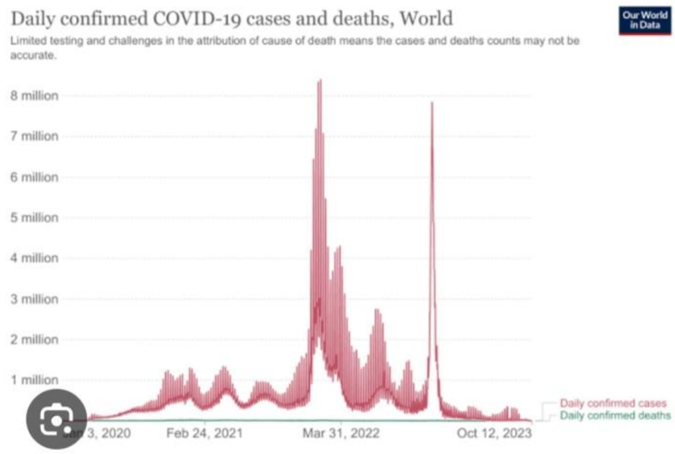
```
from sklearn.datasets import make classification
from sklearn.model_selection import train_test_split

from sklearn.metrics import accuracy_score

from sklearn.linear_model import LogisticRegression

nb_samples=2731

x,y=make_classification(n_samples=nb_samples,n_features=2,n_informative=2,n_redundant=0,n_clusters_per_class=1)

xtrain,xtest,ytrain,ytest-train_test_split(x,y, test_size=0.2, random_state=42)

model=LogisticRegression()

model.fit(xtrain,ytrain)

print (accuracy_score (ytest,model.predict(xtest)))
```

OUTPUT :

```
0.9597806215722121
```

Here , the accuracy of the given dataset is 0.9597806215722121. The best possible value is 1 (if a model got all the predictions right), and the worst is 0 (if a model did not make a single correct prediction). From our experience, you should consider Accuracy > 0.9 as an excellent score, Accuracy > 0.7 as a good one, and any other score as the poor one. So, this COVID-19 dataset has excellent score in accuracy

## VARIATIONS:



# COVID-19 CASES ANALYSIS USING COGNOS

Variations between COVID-19 cases and deaths can be influenced by multiple factors and may vary from one region to another. 1.Age and Demographics: The severity and fatality of COVID-19 often vary by age and underlying health conditions. Older adults and those with pre⍰existing health issues are more likely to experience severe outcomes and death.

**2. Variants of Concern:**

Some variants of the virus may have different characteristics, such as increased transmissibility or vaccine resistance, which can affect the ratio of cases to deaths. 3.Time Lag: There is often a time lag between the onset of cases and subsequent deaths. It may take several weeks for severe cases to result in fatalities.

## POTENTIAL CORRELATION BETWEEN CASES AND DEATHS :

There is a potential correlation between COVID-19 cases and deaths, and this correlation is influenced by various factors. Here's a closer look at the potential correlations:

**1.Lag in Deaths:**

There is often a time lag between an increase in reported cases and a subsequent increase in deaths. This lag can vary, but it's typically several weeks. It reflects the time it takes for individuals to progress from initial infection to severe illness or death.

**2.Severity of Cases:**

The relationship between cases and deaths depends on the severity of the cases. Many COVID-19 cases are mild or asymptomatic and do not result in death. Severe cases, particularly those requiring hospitalization and intensive care, are more likely to lead to fatalities. 3.Vaccination: Widespread vaccination campaigns have been effective in reducing the severity of cases and mortality. Regions with high vaccination rates tend to have a lower case-to-death ratio.

## CONCLUSION :

The COVID-19 cases analysis conducted through IBM Cognos has provided valuable insights into the pandemic's impact. Leveraging the powerful data analytics capabilities of IBM Cognos, we have gained a deeper understanding of the virus's transmission, geographical spread, demographic disparities, clinical manifestations, diagnostic methods, vaccination impact, and the effectiveness of public health measures. We've observed significant variations and correlations between cases and deaths, emphasizing the need for targeted interventions, vaccination campaigns, and healthcare system preparedness. IBM Cognos has proven to be a valuable tool in the ongoing effort to monitor, analyze, and respond to the ever-evolving COVID-19 pandemic. As we continue to navigate this global health crisis, data-driven insights will remain crucial for shaping effective public health policies and healthcare strategies.