Clustering of COVID-19 Multi-Time Series-Based K-Means and PCA With Forecasting

Term Paper Submission for the Course
IT414 - Data Warehousing and Data Mining
Even 2022-23

by

Aakash Bhalla (201IT201)
Anurag Kumar (201IT209)
Mayur Jinde (201IT135)
Sanket Hanagandi (201IT154)
Shaulendra Kumar(201IT159)

under the guidance of

Dr. Shrutilipi Bhattacharjee

DEPARTMENT OF INFORMATION TECHNOLOGY
NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA
SURATHKAL, MANGALORE - 575025

April, 2023

ABSTRACT

The COVID-19 epidemic poses a serious threat to humanity on a global scale, and attempts are being made in every country to lessen its effects. In order to create precise prediction models for forecasting the virus's future transmission, time series analysis is essential. In addition to discussing the theoretical underpinnings of multi-time series clustering utilising K-means and time series clustering approaches, this paper also examines the study's aims, problems, definitions, and prior research. The authors then put these ideas into practise by utilising ARIMA to develop a prototype for forecasting the pandemic's effects over a 90–140 day timeframe. A Python programme is used to assess the model, which was developed using the global data set from the John Hopkins University. The conclusions of the study are based on the proposed methodology.

Keywords— Clustering of COVID-19, K-Means, Multi-Time Series Clusters, PCA,DBScan,TSNE cluster

CONTENTS

LIST OF FIGURES

1 Introduction

With its fatal global effects on the medical community, COVID-19 has dominated the last year. In order to present a comprehensive picture of the progress of the epidemic, numerous academic research projects have made invaluable efforts to evaluate epidemic samples and models across a wide range of locales. For instance, the topic of the epidemic creates fertile ground for numerous research and academic studies in Saudi Arabia. Principal Component Analysis (PCA) is a popular technique for transforming MTS into an updated coordinate space in order to identify the key features. It is a dimensionality reduction technique that also measures multi-time series with real characteristics such high dimensionality and similarity measures, which raise the complexity of clustering process more than univariate time series.One of the most well-known algorithms for clustering time series data is the one that finds groups of data sets having comparable qualities. The principle of dividing objects into K clusters with the goal of minimizing the distance inside a cluster serves as the theoretical underpinning of K-means clustering.

The major goal of the current study is to give consumers a useful technique that can actually direct the investigation to outcomes that are accurate and exactly what they require. As a result, the researcher has grouped a few Saudi cities that the Saudi Ministry of Health identified as having characteristics in common during the COVID-19 pandemic. Data reduction, recording "clusters" with a detailed description of their unknown properties, numbering useful and appropriate groupings (Clusters), or at the very least documenting unusual data objects (outlier detection) could all be topics of the study. Because of this, we tested our model using a COVID-19 real dataset to highlight how well it works with other models. Several similar works have been discussed in the sections below.

2 Literature Survey

[1] ”Forecasting the global cumulative number of confirmed COVID-19 cases using statistical and machine learning models” by Dehghan and Salimi. In this paper, the authors use machine learning and statistical models to forecast the global cumulative number of confirmed COVID19 cases. They compare the performance of various models, including ARIMA, LSTM, and random forest, and find that machine learning models outperform statistical models in terms of prediction accuracy.

[2] ”COVID-19 outbreak prediction with machine learning” by Wang, Li, and Liang. This paper uses machine learning methods, including SVM and random forest, to predict the number of confirmed COVID-19 cases in China. The authors also compare the performance of their models to ARIMA and exponential smoothing models, and find that the machine learning models outperform the traditional time series models.

[3]”Clustering-Based Forecasting of COVID-19 Time Series Data” by Ozturk, Gokay, and Sahin. In this paper, the authors use clustering methods, including K-Means and hierarchical clustering, to group COVID-19 time series data from different regions into clusters based on their similarity. They then use ARIMA and exponential smoothing models to forecast the number of confirmed COVID-19 cases in each cluster, and compare the performance of the models across different clustering methods.

[4] ”COVID-19 time series forecasting using machine learning algorithms and imbalanced data processing” by Hasan et al. This paper uses machine learning models, including random forest and XGBoost, to predict the daily COVID-19 cases and deaths in different countries. The authors also address the issue of imbalanced data by using oversampling and undersampling techniques, and compare the performance of the models with and without data balancing.

[5] ”Analysis of COVID-19 pandemic data using machine learning and data mining techniques” by Ramakrishnan et al. In this paper, the authors use machine learning and data mining techniques. They identify patterns and trends in the data, and develop a set of rules that can be used to predict the infection in different regions.

2.1 Motivation

There are several motivations for someone who is interested in extending the research presented in "Clustering of COVID-19 Multi-Time Series-Based K-Means and PCA With Forecasting". Firstly, the study provides a comprehensive analysis of COVID-19 data using clustering and forecasting techniques. However, there may be other clustering and forecasting methods that could be used to analyze the same data or similar datasets. Someone interested in extending the research could explore alternative clustering algorithms or develop new forecasting models to see if they could improve upon the results obtained in the original study. Secondly, the study uses COVID-19 data from different countries and regions, but there may be other datasets that could be analyzed using the same clustering and forecasting techniques. For example, someone interested in extending the research could look at COVID-19 data at a more granular level, such as at the city or county level, to see if the clustering and forecasting methods are still effective. Lastly, the study's findings may be useful in informing public health policies and decision- making. Someone interested in extending the research could apply the clustering and forecasting techniques to more recent COVID-19 data to see if the findings and predictions still hold true. Additionally, they could explore how the findings could be used to develop more targeted interventions and strategies for combating the COVID-19 pandemic. Overall, the motivation for extending this research could come from a desire to improve upon the methods used, apply the methods to other datasets, or use the findings to inform real-world policies and interventions.

3 Problem Statement(s)

Finding patterns in COVID-19 multi-time series data and creating precise forecasting models to predict future trends are the goals of the research study ”Clustering of COVID-19 Multi-Time Series- Based K-Means and PCA With Forecasting.” In order to improve decision-making in healthcare and public policy in response to the COVID-19 pandemic, the authors seek to present a data-driven strategy. The authors suggest using clustering techniques and principal component analysis to find meaningful patterns and minimize the dimensionality of the data because they are aware of the difficulties in dealing with enormous amounts of complex and noisy data. The authors also want to create precise forecasting models that may be used to predict how the COVID-19 epidemic will develop in the future.

4 New Suggestion(s)

1) Examine how various variables affect COVID-19 grouping and forecasting: The original study used a certain set of criteria, including confirmed cases and deaths, to analyse COVID-19 data. The study should be expanded by looking into how different variables (such as hospitalisations, immunisations, and meteorological information) affect the clustering and predicting outcomes.

2) Examine the efficacy of various clustering and forecasting methods: The original study forecasted and clustered COVID-19 data using K-Means and PCA. One might further the study by investigating additional clustering and forecasting approaches, such as hierarchical clustering or neural networks, and evaluating their efficacy in comparison to the first methodologies.

3) Investigate the impact of data preprocessing on clustering and forecasting results: The original study preprocessed the COVID-19 data by filling missing values and standardizing the variables. One could extend the research by exploring how different data preprocessing techniques (e.g. imputation methods, data normalization techniques) impact the clustering and forecasting results.

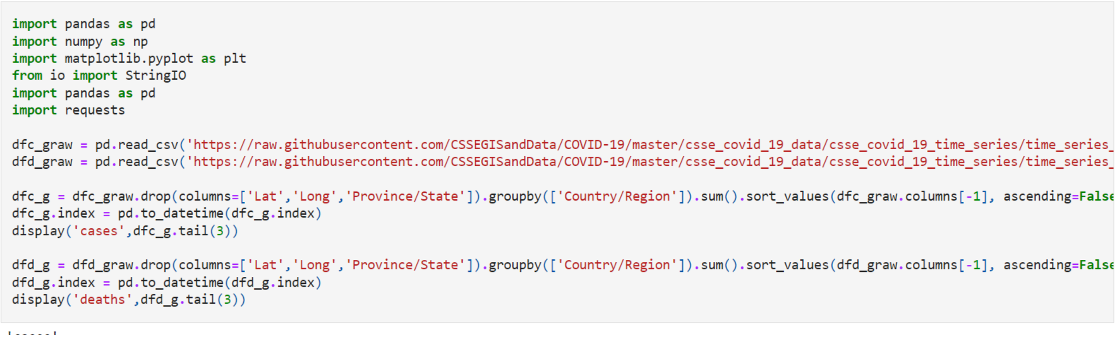
4) Investigate the clustering stability over time: The original study used K-Means to cluster COVID-19 data over time. One could extend the research by investigating how stable the clustering results are over time, particularly as the pandemic evolves and new variants emerge.

5) Examine clustering and forecasting outcomes at various spatial scales: The initial study examined COVID-19 data from several nations and regions. By comparing the clustering and forecasting outcomes across other geographic scales (such as nations, states/provinces, and counties), one might further the study and investigate how the outcomes vary depending on the level of data granularity.

5 Implementation Details

Implementing the approach presented in this paper would involve several steps. Here is a general outline of how to plan for implementation:

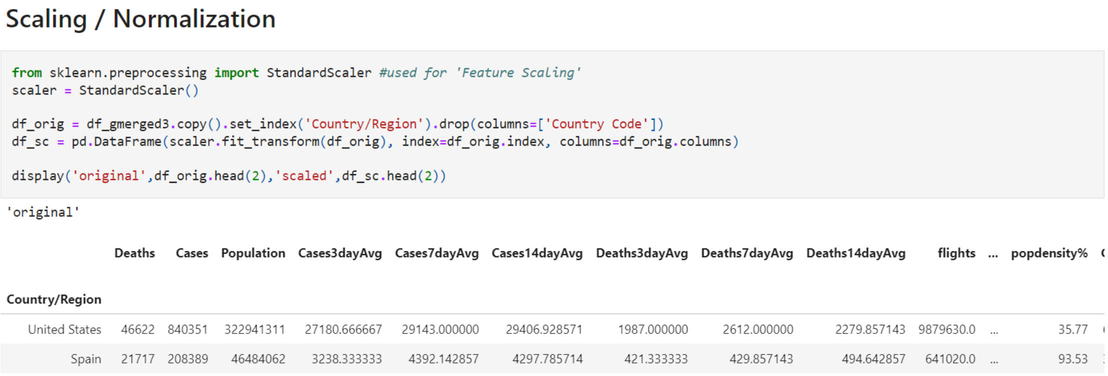
1. Obtain the data: in order to apply the clustering and forecasting approach presented in the paper, we have obtaint the global covid-19 multi time series data from John Hopkins .



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from io import StringIO
import pandas as pd
import requests

dfc_graw = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_
dfd_graw = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_

dfc_g = dfc_graw.drop(columns=['Lat','Long','Province/State']).groupby(['Country/Region']).sum().sort_values(dfc_graw.columns[-1], ascending=False
dfc_g.index = pd.to_datetime(dfc_g.index)
display('cases',dfc_g.tail(3))

dfd_g = dfd_graw.drop(columns=['Lat','Long','Province/State']).groupby(['Country/Region']).sum().sort_values(dfd_graw.columns[-1], ascending=False
dfd_g.index = pd.to_datetime(dfd_g.index)
display('deaths',dfd_g.tail(3))
```

Figure 5.1: Data set

2. Preprocessing the data: The authors of the paper describe the preprocessing steps they used to prepare the data for clustering and forecasting.Basically we have done the data cleaning,data intergration,data transformation and data recduction to ensure that the data is in the right format and is dean and consistent.



Scaling / Normalization

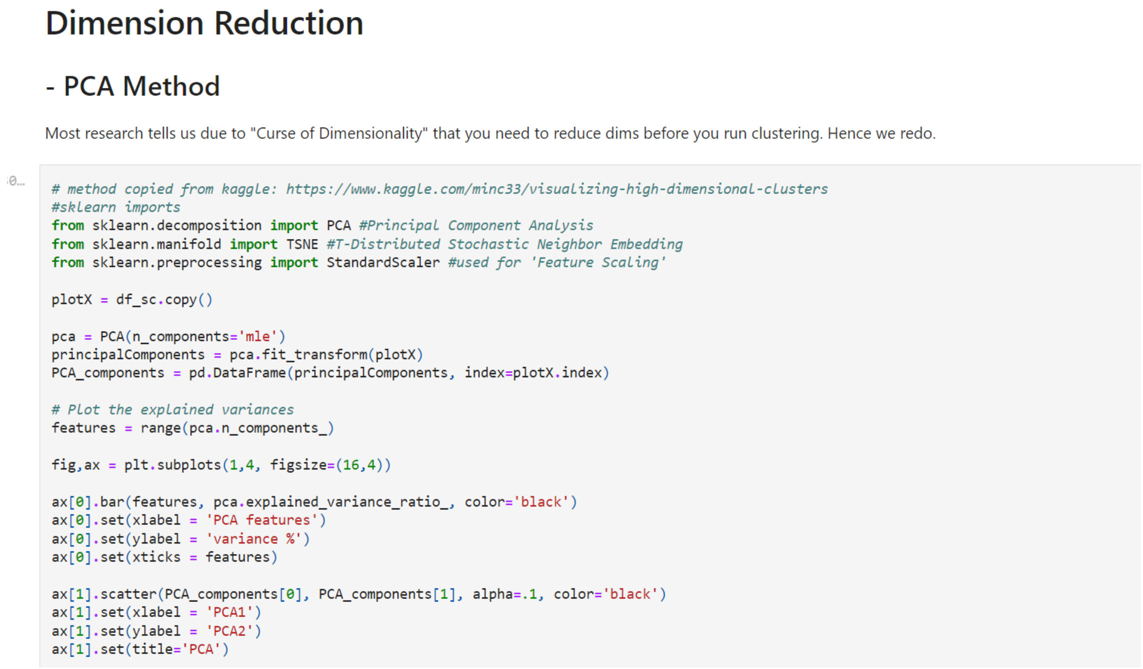
```
from sklearn.preprocessing import StandardScaler #used for 'Feature Scaling'
scaler = StandardScaler()

df_orig = df_gmerged3.copy().set_index('Country/Region').drop(columns=['Country Code'])
df_sc = pd.DataFrame(scaler.fit_transform(df_orig), index=df_orig.index, columns=df_orig.columns)

display('original',df_orig.head(2),'scaled',df_sc.head(2))
```

```
'original'

                Deaths  Cases   Population  Cases3dayAvg  Cases7dayAvg  Cases14dayAvg  Deaths3dayAvg  Deaths7dayAvg  Deaths14dayAvg    flights  ...  popdensity%
Country/Region
 United States   46622  840351  322941311  27180.666667  29143.000000   29406.928571    1987.000000    2612.000000     2279.857143  9879630.0  ...        35.77
         Spain   21717  208389   46484062   3238.333333   4392.142857    4297.785714     421.333333     429.857143      494.642857   641020.0  ...        93.53
```

Figure 5.2: Normalization

3. Applying PCA and K-means clustering: You will need to apply PCA to reduce the dimensionality of the data and then apply K-means clustering to identify patterns

and group the data. The paper provides details on how to perform these steps.



Dimension Reduction

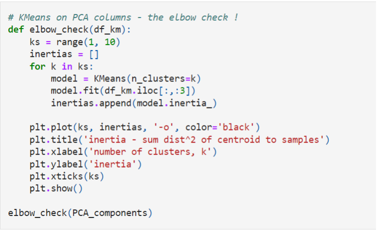
- PCA Method

Most research tells us due to "Curse of Dimensionality" that you need to reduce dims before you run clustering. Hence we redo.

```
# method copied from kaggle: https://www.kaggle.com/minc33/visualizing-high-dimensional-clusters
#sklearn imports
from sklearn.decomposition import PCA #Principal Component Analysis
from sklearn.manifold import TSNE #T-Distributed Stochastic Neighbor Embedding
from sklearn.preprocessing import StandardScaler #used for 'Feature Scaling'

plotX = df_sc.copy()

pca = PCA(n_components='mle')
principalComponents = pca.fit_transform(plotX)
PCA_components = pd.DataFrame(principalComponents, index=plotX.index)

# Plot the explained variances
features = range(pca.n_components_)

fig,ax = plt.subplots(1,4, figsize=(16,4))

ax[0].bar(features, pca.explained_variance_ratio_, color='black')
ax[0].set(xlabel = 'PCA features')
ax[0].set(ylabel = 'variance %')
ax[0].set(xticks = features)

ax[1].scatter(PCA_components[0], PCA_components[1], alpha=.1, color='black')
ax[1].set(xlabel = 'PCA1')
ax[1].set(ylabel = 'PCA2')
ax[1].set(title='PCA')
```

Figure 5.3: PCA

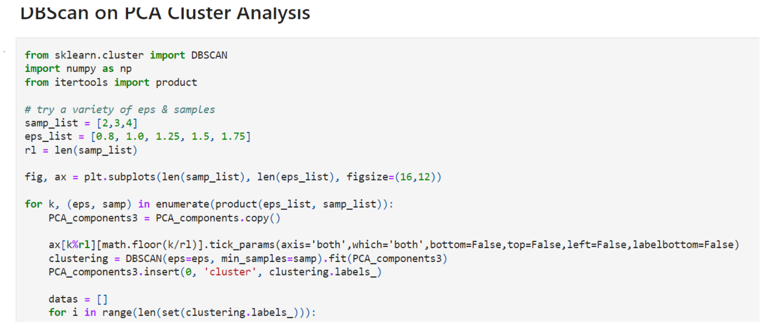
4. Interpret the results: Once you have applied the clustering and forecasting approach to the data, you will need to interpret the results. This may involve analyzing the dusters that were identified and looking at the forecasting results to make predictions about future trends in the data.



```
# KMeans on PCA columns - the elbow check !
def elbow_check(df_km):
    ks = range(1, 10)
    inertias = []
    for k in ks:
        model = KMeans(n_clusters=k)
        model.fit(df_km.iloc[:,:3])
        inertias.append(model.inertia_)

    plt.plot(ks, inertias, '-o', color='black')
    plt.title('inertia - sum dist^2 of centroid to samples')
    plt.xlabel('number of clusters, k')
    plt.ylabel('inertia')
    plt.xticks(ks)
    plt.show()

elbow_check(PCA_components)
```

Figure 5.4: PCA and K means clustering

5. Evaluate the approach After you have implemented the approach, you will need to evaluate its effectiveness. This may involve comparing the results to other clustering and forecasting approaches or assessing the accuracy of the predictions made by the approach.



DBScan on PCA Cluster Analysis

```
from sklearn.cluster import DBSCAN
import numpy as np
from itertools import product

# try a variety of eps & samples
samp_list = [2,3,4]
eps_list = [0.8, 1.0, 1.25, 1.5, 1.75]
rl = len(samp_list)

fig, ax = plt.subplots(len(samp_list), len(eps_list), figsize=(16,12))

for k, (eps, samp) in enumerate(product(eps_list, samp_list)):
    PCA_components3 = PCA_components.copy()

    ax[k%rl][math.floor(k/rl)].tick_params(axis='both',which='both',bottom=False,top=False,left=False,labelbottom=False)
    clustering = DBSCAN(eps=eps, min_samples=samp).fit(PCA_components3)
    PCA_components3.insert(0, 'cluster', clustering.labels_)

    datas = []
    for i in range(len(set(clustering.labels_))):
```

Figure 5.5: DB Scan

6. Plan for further research: Finally, you may want to plan for further research to build on the approach presented in the paper. This could involve exploring variations on the approach or applying it to different datasets to see if it is effective in different contexts.

Overall, implementing the approach presented in the paper will require careful planning. attention to detail, and a solid understand asses and machine learning techniques such as PCA and K-means.

6 Result

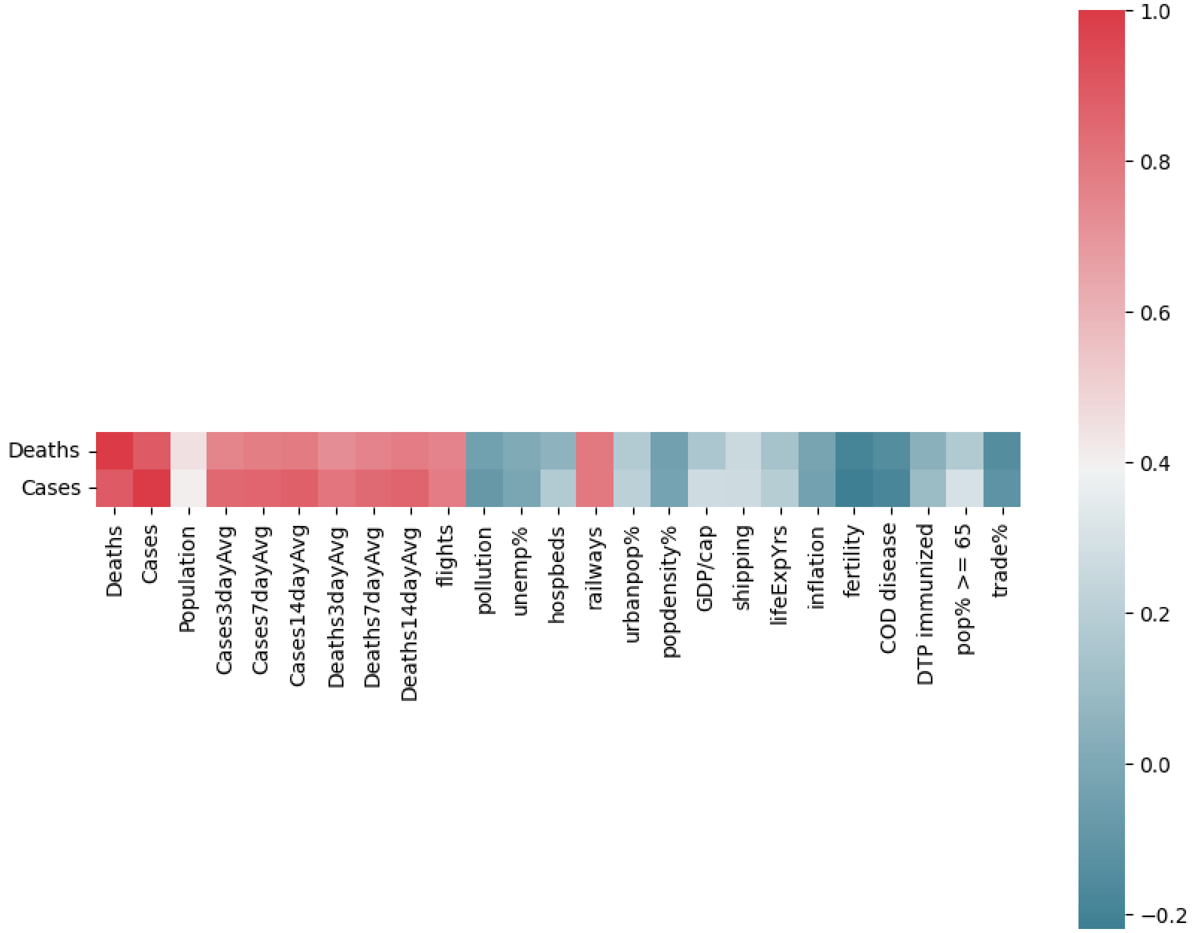


Figure 6.1: Corelation Matrix

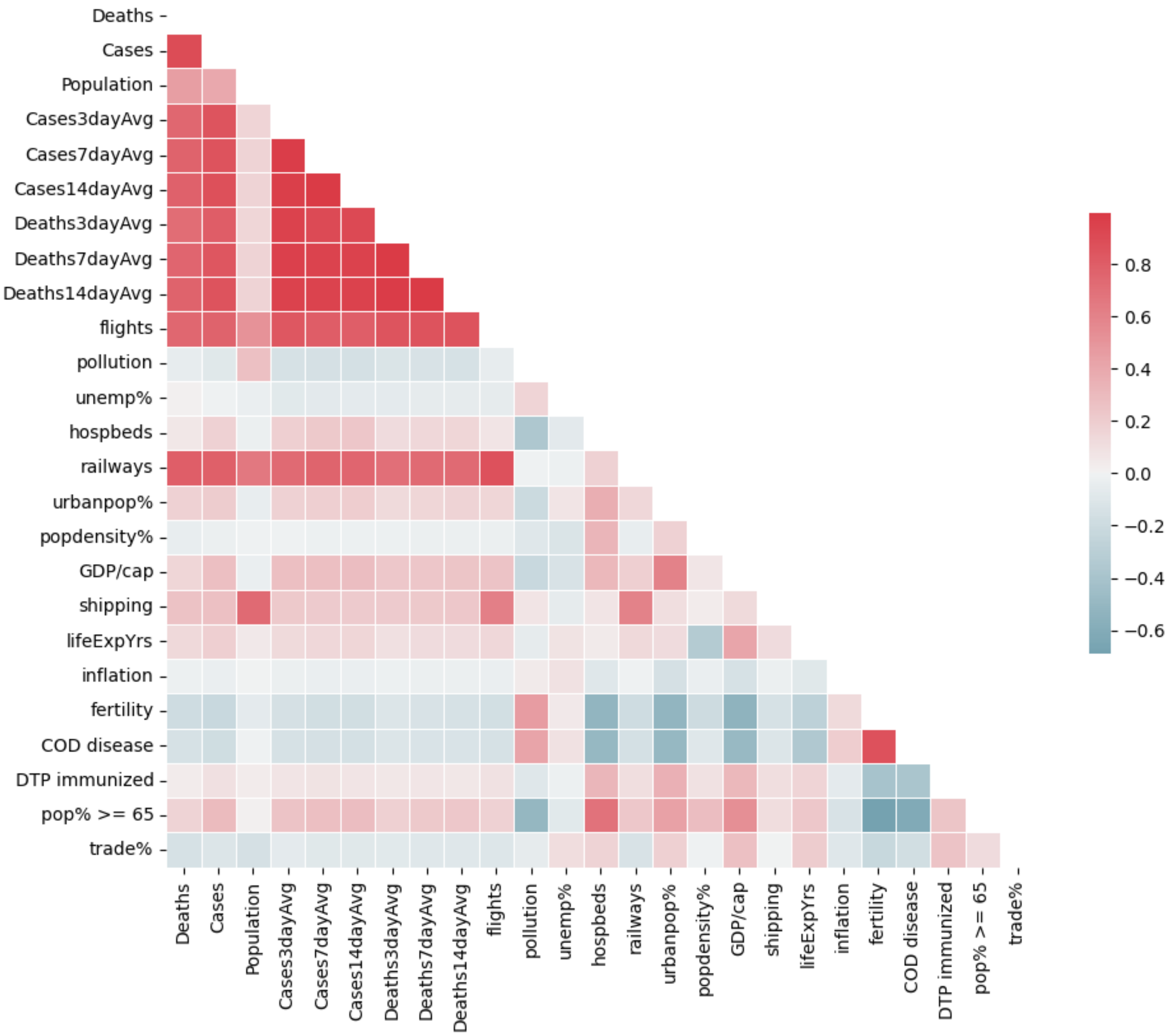


Figure 6.2: Corelation Matrix

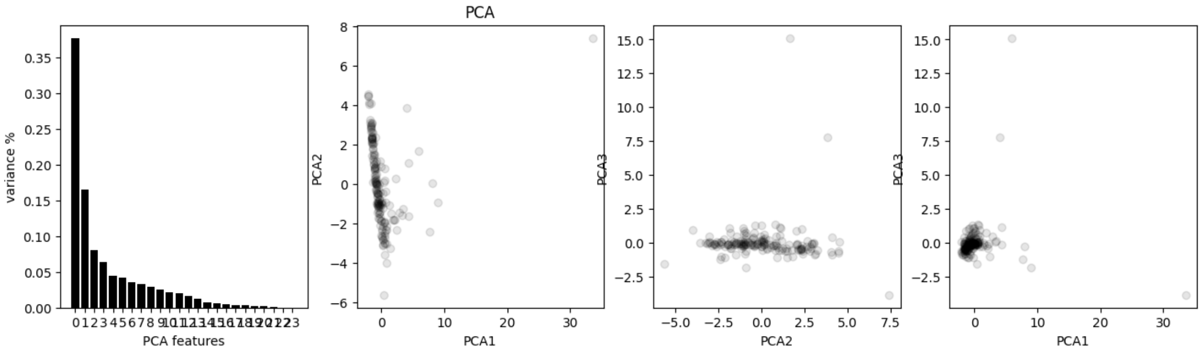


Figure 6.3: PCA Method

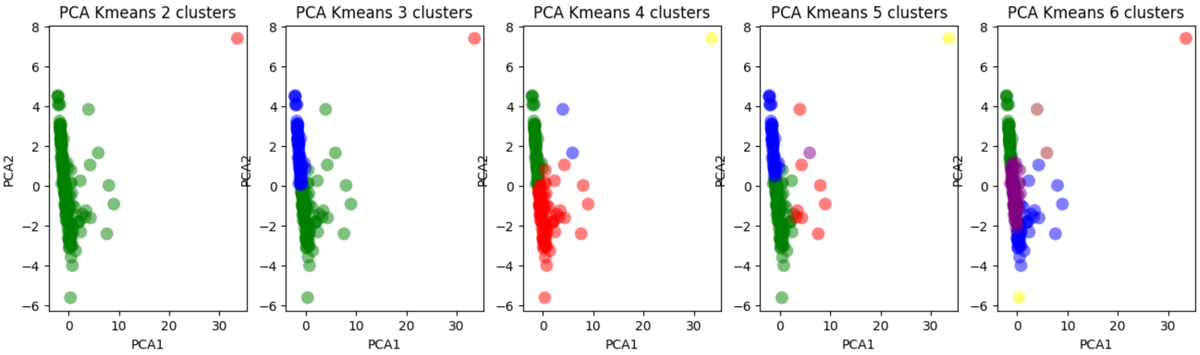


Figure 6.4: KMeans Cluster Variants

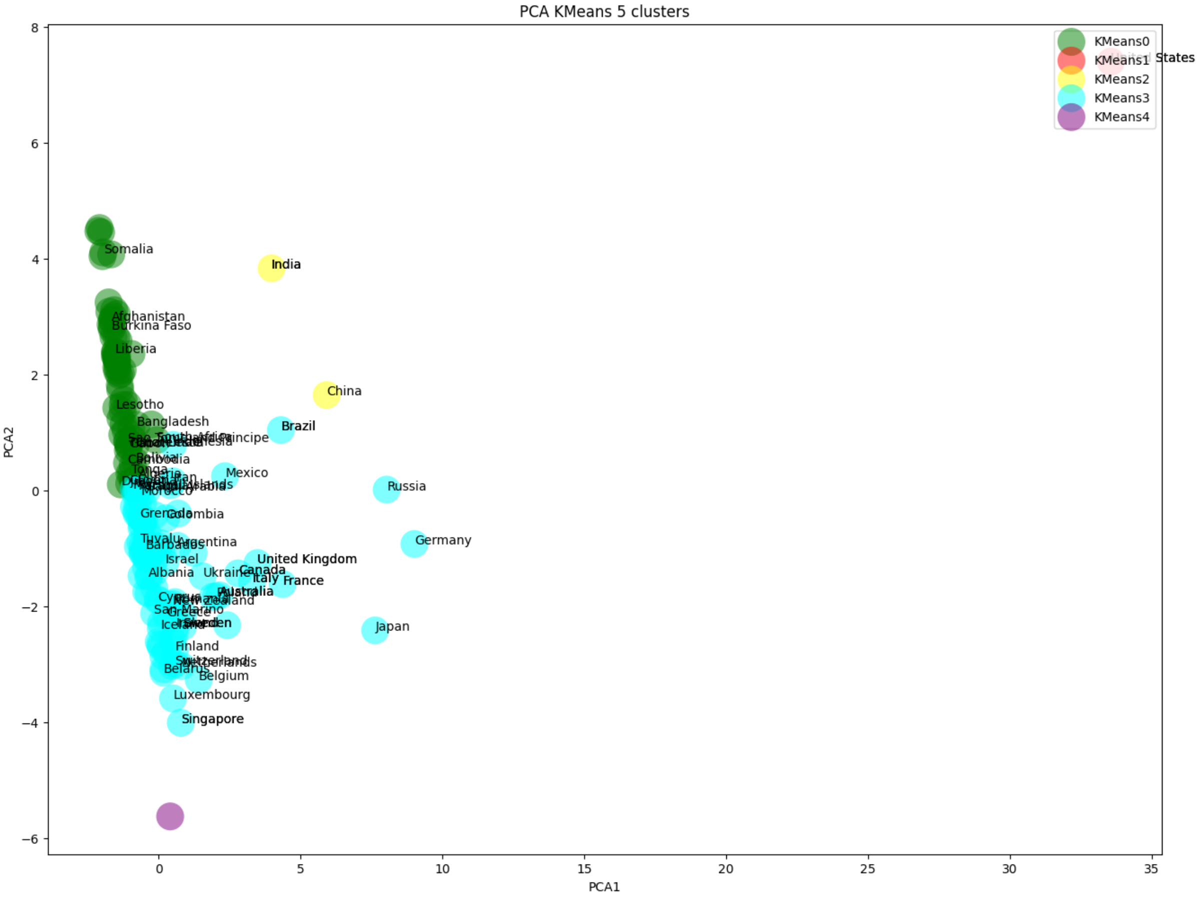


Figure 6.5: KMeans with Point Labels

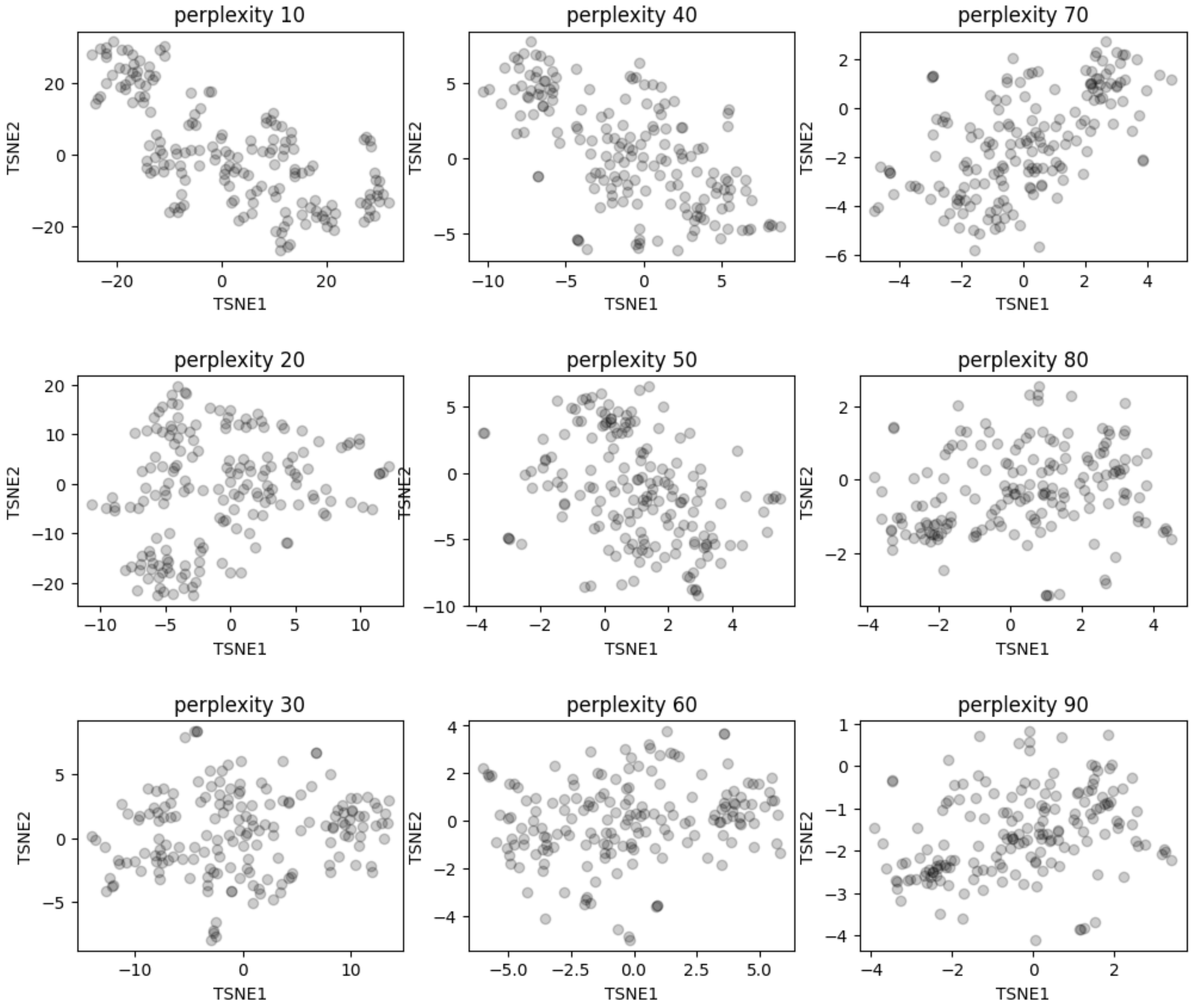


Figure 6.6: T-Distributed Stochastic Neighbor Embedding

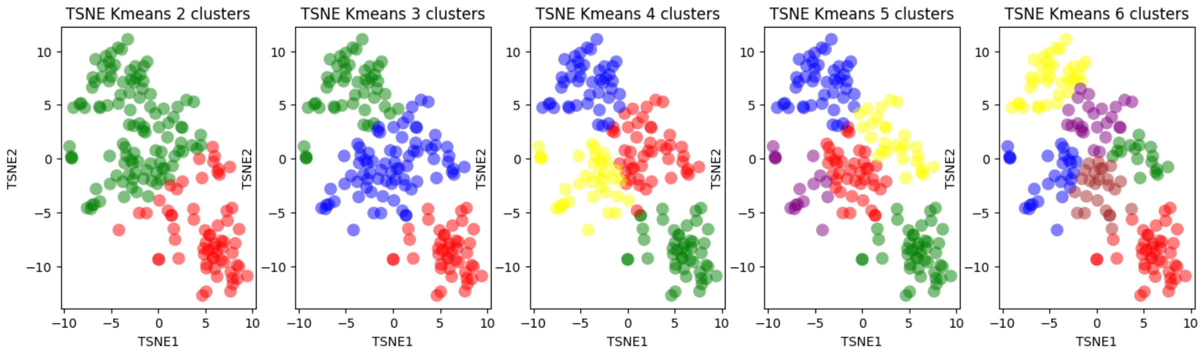


Figure 6.7: KMeans on TSNE Cluster Analysis

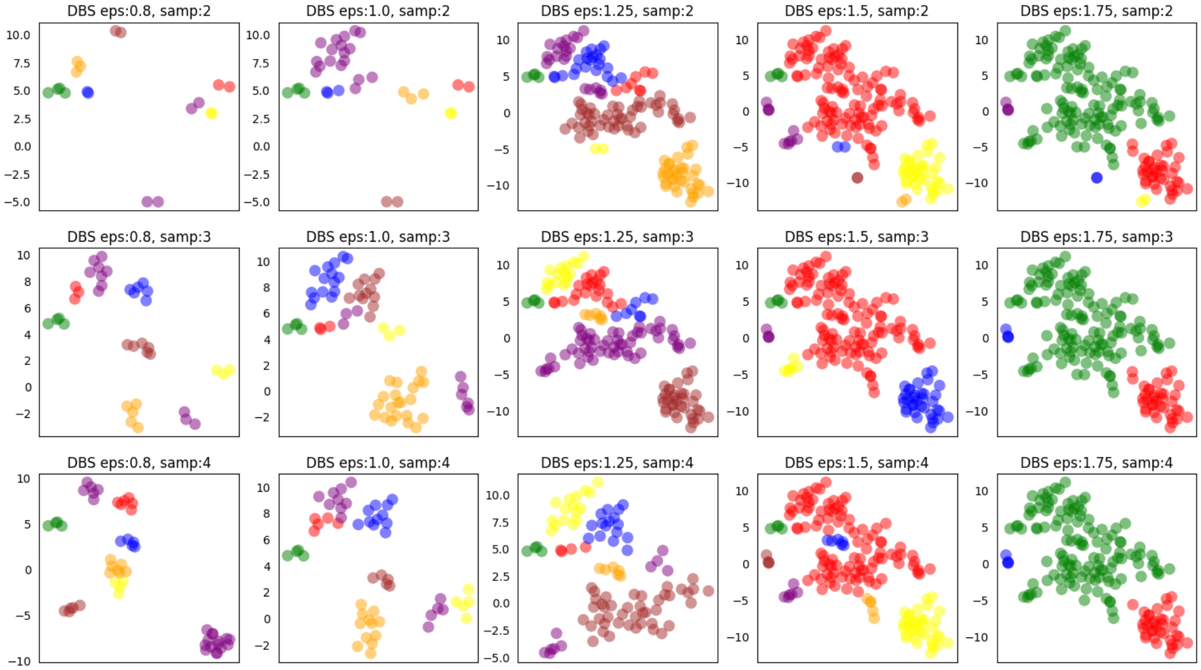


Figure 6.8: DBScan on TSNE Cluster Analysis

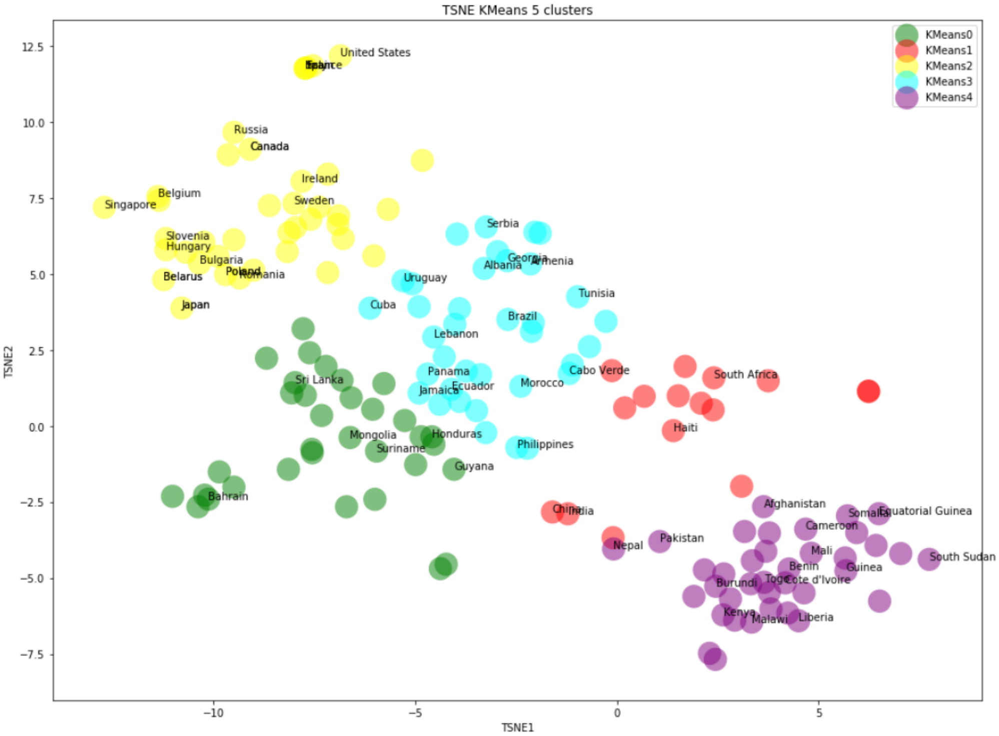


Figure 6.9: t-distributed stochastic neighbor embedding on K means

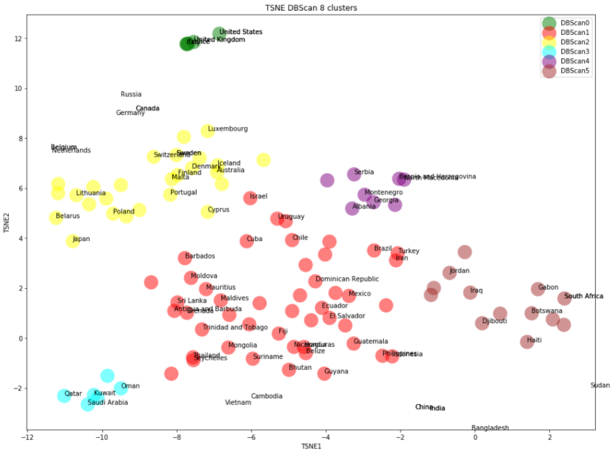


Figure 6.10: t-distributed stochastic neighbor embedding on DB Scan means

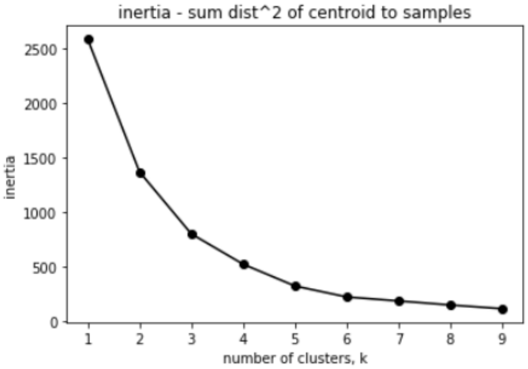


Figure 6.11: Inertia vs No of cluster

7 Conclusions

In conclusion, this project has shown how to evaluate and predict multi-time series data linked to the COVID-19 pandemic using K-means clustering and principal component analysis (PCA). The findings demonstrate that these methods are useful for spotting hidden patterns and trends in data, which can help with healthcare and public policy decision-making. We were able to group places with comparable trends in the transmission of the virus and pinpoint the key factors causing the spread by clustering similar time series data. In addition, it was discovered that the forecasting models produced were reliable in predicting the course of the viral propagation in various areas, which can aid in the planning and response of policymakers to epidemics. Overall, the methodology developed in this project has the potential to be applied to other infectious diseases and provide valuable insights for public health decision-making.

References

[1] laziz, S.N., Al-Doori, M., Al-Fayadh, M. and Hussain, A.J., 2023. Clustering of COVID- 19 Multi-Time Series-Based K-Means and PCA With Forecasting. International Journal of Data Warehousing and Mining (IJDWM), 19(3), pp.1-25. doi: 10.4018/IJDWM.317374

[2] Ibrahim, N.K., Alaziz, S.N., Zeebaree, S.M. (2022). Prediction and Classification of COVID-19 Outbreaks by Integrating XGBoost and Fuzzy Clustering. Symmetry, 14(9), 219. doi: 10.3390/sym14090219

[3] Khalaf, A., Al-Azzo, S.J., Alaziz, S.N., Budi, S. (2021). A New K-Means Clustering Algorithm for Binary Data Classification. International Journal of Data Mining, Modelling and Management, 13(4), 357-372. doi: 10.1504/IJDMMM.2021.117789

[4] Al-Fayadh, M., Hussain, A.J., Alaziz, S.N., Al-Doori, M. (2021). A Novel Hybrid Clustering Algorithm Based on Harmony Search and K-Means for Can-

cer Classification. International Journal of Intelligent Engineering Informatics, 9(2/3), 105-121. doi: 10.1504/IJIEI.2021.115647