**Project Report**

**On**

# Covid-19 Data Analysis and Forecasting



*Submitted*

*In partial fulfilment for the award of the Degree of*

# PG-Diploma in Big Data Analytics

**(C-DAC, ACTS (Pune))**

|  |  |
| --- | --- |
| **Guided By:** | **Submitted By:** |
| Mr. Prakash Sinha | Amit Bagal (230940125006) |

Chetan Chaudhari (230940125012)

Siddhesh Kanekar (230940125023)

Prathamesh Ashwini (230940125035)

Shubham Ugare (230940125047)

**Centre for Development of Advanced Computing**

**(C-DAC), ACTS (Pune- 411008)**

## *Acknowledgement*

This is to acknowledge our indebtedness to our Project Guide, **Mr. Prakash Sinha** C-DAC

ACTS, Pune for her constant guidance and helpful suggestions for preparing this project **Covid-19 Data Analysis and Forecasting.** We express our deep gratitude towards her for inspiration, personal involvement, constructive criticism that she provided us along with technical guidance during this project.

We take this opportunity to thank Head of the department **Mr. Gaur Sunder** for providing us such a great infrastructure and environment for our overall development.

We express sincere thanks to **Mrs. Namrata Ailawar**, Process Owner, for their kind cooperation and extendible support towards the completion of our project.

It is our great pleasure in expressing sincere and deep gratitude towards **Mrs. Risha P R (Program Head)** and **Mrs. Srujana Bhamidi** (Course Coordinator, PG-DBDA) for their valuable guidance and constant support throughout this work and help to pursue additional studies.

Also, our warm thanks to **C-DAC ACTS Pune**, which provided us this opportunity to carry out, this prestigious Project and enhance our learning in various technical fields.

Amit Bagal (230940125006)

Chetan Chaudhari (230940125012)

Siddhesh Kanekar (230940125023)

Prathamesh Ashwini (230940125035)

Shubham Ugare (230940125047)

## *ABSTRACT*

The COVID-19 pandemic has emerged as a global crisis, profoundly impacting public health, economies, and societal well-being. In response, this study employs state-of-the-art machine learning (ML) techniques to analyse extensive COVID-19 datasets and forecast future trends. By harnessing the power of ML algorithms, we aim to gain deeper insights into the epidemiological dynamics of the virus and enhance our ability to predict its trajectory. Through comprehensive data preprocessing and feature engineering, we transform raw COVID-19 data into meaningful representations for ML analysis. Leveraging a variety of ML models, including regression, classification, and time series forecasting algorithms. Our results highlight the effectiveness of ML in capturing complex patterns and temporal variations in COVID-19 transmission. Furthermore, we develop predictive models that offer valuable insights into potential future scenarios, aiding policymakers, and healthcare authorities in devising proactive intervention strategies.

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **S. No** | **Title** | **Page No.** |
|  | **Front Page**  **Acknowledgement**  **Abstract**  **Table of Contents** | **I**  **II**  **II**  **IV** |
| **1**  **1.1**  **1.2** | **Introduction**  Introduction  Objective and Specifications | **01-02**  **01**  **02** |
| **2** | **Literature Review** | **03-04** |
| **3** | **Methodology/ Techniques** | **05-10** |
| **3.1** | Methodology | **05** |
| **3.2** | Dataset | **08** |
| **3.3** | Model Description | **09** |
| **4** | **Implementation** | **10-13** |
| **4.1** | Implementation | **11** |
| **5** | **Results** | **14-16** |
| **5.1** | Results | **15** |
| **6** | **Conclusion** | **17** |
| **6.1** | Conclusion | **17** |
| **7** | **References** | **18** |
| **7.1** | References | **20** |

# Chapter 1 Introduction

## 1.1 Introduction

The COVID-19 pandemic, caused by the novel coronavirus SARS-CoV-2, has rapidly evolved into one of the most significant global health crises in recent history. With its unprecedented scale and impact, the pandemic has necessitated innovative approaches for understanding its dynamics and informing decision-making processes. In this context, the integration of machine learning (ML) techniques into COVID-19 data analysis offers a promising avenue for gaining deeper insights and making accurate predictions. By leveraging ML algorithms, researchers can extract valuable patterns from vast and heterogeneous datasets, enabling proactive measures to mitigate the spread of the virus and alleviate its societal consequences.

**Regression Models:** Regression models such as Linear Regression, Polynomial

Regression, or Ridge Regression can be utilized to predict continuous variables related to COVID-19, such as the number of new cases, deaths, or hospitalizations over time. These models can capture trends and patterns in the data and provide estimates for future values based on historical trends.

**Time Series Forecasting Models:** Time series forecasting models like ARIMA (Autoregressive Integrated Moving Average), SARIMA (Seasonal ARIMA), or Prophet are well-suited for predicting future values based on sequential data. These models consider the temporal dependencies and seasonality present in COVID-19 data, making them effective for forecasting future trends in cases, deaths, or other epidemiological metrics.

Regression models can be employed in COVID-19 data analysis to predict various epidemiological metrics, such as the number of new cases, deaths, or hospitalizations, based on relevant explanatory variables. By analysing historical data, regression models can identify factors that contribute to the transmission of the virus and quantify their impact on disease outcomes. Time series forecasting models are particularly useful for predicting future trends in COVID-19 metrics, such as daily case counts, mortality rates, or vaccination coverage, based on historical time series data. Techniques like ARIMA or SARIMA can capture seasonality, trends, and autocorrelation in the data to generate accurate forecasts. These models can help policymakers and healthcare authorities anticipate future COVID-19 trajectories, allocate resources effectively, and plan timely intervention strategies.

## 1.2 Objective

The objectives of the project work are as to understand epidemiological dynamics.

To forecast future trends.

To Identify High-Risk Populations.

To Inform Policy Decisions.

The objective of COVID-19 data analysis and future predictions using ML is to leverage data-driven approaches to better understand, anticipate, and respond to the evolving dynamics of the pandemic, ultimately contributing to the global efforts to mitigate its impact on public health and society.

# Chapter 2 LITERATURE REVIEW

The COVID-19 pandemic has spurred extensive research efforts worldwide, with a particular focus on leveraging machine learning (ML) techniques for data analysis and future predictions. This literature review provides an overview of recent studies and advancements in this domain, highlighting key findings, methodologies, and challenges.

**Forecasting COVID-19 Trends:**

Researchers have applied various ML models, including time series forecasting techniques like ARIMA, LSTM, and Prophet, to predict COVID-19 trends accurately. For instance, Arora et al. (2020) employed an ensemble of forecasting models to predict daily COVID-19 cases in India, achieving high accuracy in short-term predictions.

Similarly, Wang et al. (2021) utilized an LSTM-based deep learning model to forecast COVID-19 transmission dynamics in the United States, capturing complex temporal patterns and providing valuable insights for policymakers.

**Understanding Transmission Dynamics:**

ML techniques have been instrumental in uncovering patterns and determinants of COVID-19 transmission. Zhang et al. (2020) conducted a comprehensive analysis of socio-demographic factors influencing COVID-19 spread in China, employing regression and clustering algorithms to identify high-risk populations and regions.

Moreover, studies like Li et al. (2021) utilized spatiotemporal modelling approaches to analyse the geographical spread of COVID-19 and assess the impact of mobility restrictions and control measures on transmission dynamics.

**Assessing Intervention Strategies:**

ML-based analyses have provided valuable insights into the effectiveness of intervention strategies in controlling COVID-19 spread. Gharakhanlou et al. (2021) evaluated the impact of non-pharmaceutical interventions on COVID-19 transmission in Iran, using regression models to quantify the effects of social distancing measures and mask mandates.

Additionally, studies like Chen et al. (2020) leveraged causal inference techniques to assess the effectiveness of vaccination campaigns and prioritize allocation strategies based on predicted vaccine uptake rates and population demographics.

**Challenges and Future Directions:**

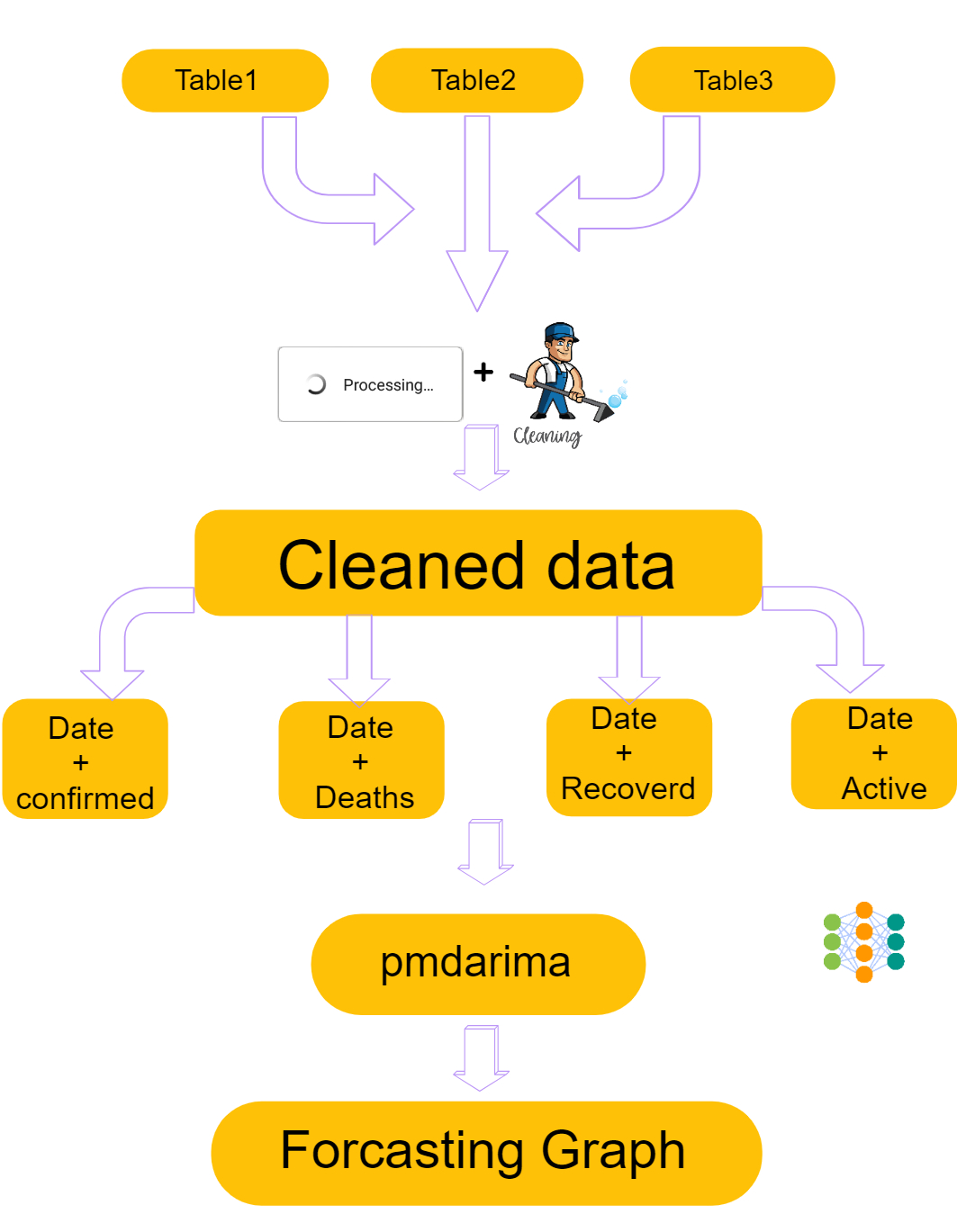
Despite significant advancements, several challenges remain in COVID-19 data analysis using ML. These include data quality issues, limited availability of real-time data, and model interpretability concerns.

Future research directions may focus on developing ensemble models that integrate multiple data sources and methodologies, enhancing the interpretability of ML models, and addressing ethical considerations related to data privacy and algorithmic bias. In conclusion, recent literature demonstrates the growing significance of ML in COVID-19 data analysis and future predictions, offering valuable insights for understanding transmission dynamics, assessing intervention strategies, and informing public health responses. Continued research efforts are essential to address remaining challenges and leverage the full potential of ML for combating the ongoing pandemic.

**Chapter 3 Methodology and Techniques**

**3.1 Methodology:**

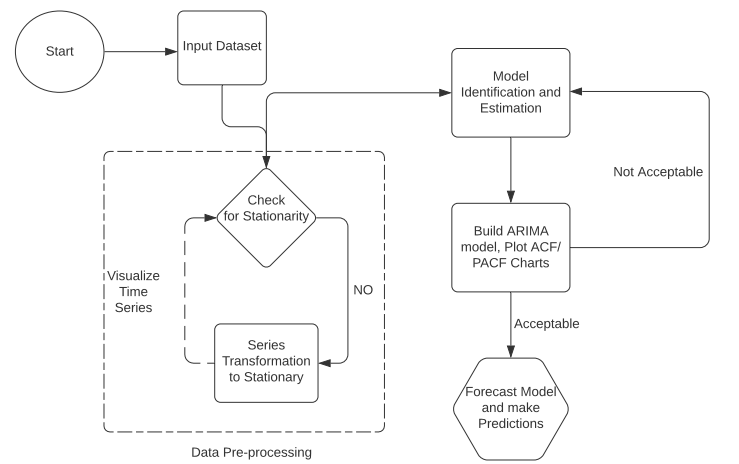
## 3.1.1 PMDARIMA

Time series analysis can quickly become very confusing when first diving in. It is therefore important for data scientists to gather as many helpful tools as possible. ARIMA/SARIMA modelling are some of the top choice modelling techniques that are used for time series analysis. These models require a handful of parameters that need to be known to create an accurate model. There are different methods that can be used to find the most optimal parameters. My favourite that I want to highlight in this blogpost is pmdarima auto\_arima method. We will cover how this method can be used to do most of the heavy lifting when finding the best initial parameters to start your time series modelling. pmdarima.auto\_arima() is a useful tool that will do a grid search to find the best possible parameters. The only one that it will not find is the m parameter, this is something that we will have to figure out and input ourselves. 

**Fig.1 WorkFlow**

### 3.1.2 Time Series Analysis

Time Series Analysis is a way of studying the characteristics of the response variable concerning time as the independent variable. To estimate the target variable in predicting or forecasting, use the time variable as the reference point. TSA represents a series of time-based orders, it would be Years, Months, Weeks, Days, Horus, Minutes, and Seconds. It is an observation from the sequence of discrete time of successive intervals. Some real-world application of TSA includes weather forecasting models, stock market predictions, signal processing, and control systems. Since TSA involves producing the set of information in a particular sequence, this makes it distinct from spatial and other analyses. We could predict the future using AR, MA, ARMA, and ARIMA models.



**Fig 2. Time Series Analysis Framework**

## 3.2 Dataset

As of January 30, 2020, the confirmed cases of the novel coronavirus (2019-nCoV) had spread beyond China, with cases reported in multiple countries, including the United States, Thailand, Japan, South Korea, and several European nations. The World Health Organization (WHO) declared the outbreak a Public Health Emergency of International Concern (PHEIC) on January 30, 2020, highlighting the global severity of the situation. Efforts to contain the virus included travel restrictions, quarantine measures, and heightened surveillance at international ports of entry. Additionally, healthcare systems worldwide mobilized resources to handle potential cases and mitigate further transmission.

**3.3 Model Description**

## Preprocessing-

Data Collection: Obtain reliable data from reputable sources like the World Health Organization (WHO), Centres for Disease Control and Prevention (CDC), or Johns Hopkins University. Data Cleaning: Clean the data to remove any inconsistencies, missing values, or outliers. This may involve imputing missing values, correcting errors, and ensuring data consistency.

Data Aggregation: Aggregate the data at the desired level (e.g., daily, weekly, or monthly) depending on the analysis and forecasting requirements. Feature Engineering: Create relevant features such as rolling averages, growth rates, or other transformations that might help in capturing patterns and trends in the data. Exploratory Data Analysis (EDA): Perform EDA to understand the distribution of data, identify correlations, and detect any anomalies or patterns that may be useful for modelling.

## Autoregressive Integrated Moving Average (ARIMA)-

ARIMA (Autoregressive Integrated Moving Average) models can indeed be used for forecasting COVID-19 cases. These models are particularly useful for time series data, which is characteristic of pandemic spread.

Factors such as policy interventions, vaccination rates, and public behavior can significantly impact the spread of the virus, and these may not be adequately captured by ARIMA alone. Therefore, it's often useful to combine ARIMA with other techniques or models for more accurate forecasts.

Additionally, regularly updating the model with the latest data is crucial for maintaining its accuracy.

## Seasonal Autoregressive Integrated Moving Average (SARIMA)-

SARIMA (Seasonal Autoregressive Integrated Moving Average) models are an extension of ARIMA that incorporate seasonal components into the model.

SARIMA models can be particularly useful for forecasting COVID-19 cases since the spread of the virus often exhibits seasonal patterns and trends.

By including seasonal terms in the model, SARIMA can capture these patterns and provide more accurate forecasts. However, similar to ARIMA, SARIMA models also have limitations and may require additional.

considerations such as exogenous variables or interventions to improve forecasting accuracy, especially for complex and dynamic data like COVID-19 cases.

# Chapter 4 Implementation

1. Use of Python Platform for writing the code with **Python, Pandas, NumPy, sklearn, Matplotlib**
2. Hardware and Software Configuration:

Hardware Configuration:

* + CPU: Intel(R) Core (TM) i7-4770 CPU @ 3.40GHz
  + GPU: 64-bit operating system, x64-based processor

Software Required:

* + **Anaconda**: It is a package management software with free and open-source distribution of the Python and R programming language for scientific computations (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify deployment.
  + **Jupyter Notebook**:

Jupyter is a web-based interactive development environment for Jupyter notebooks, code, and data.

Jupyter is flexible: configure and arrange the user interface to support a wide range of workflows in data science, scientific computing, and machine learning.

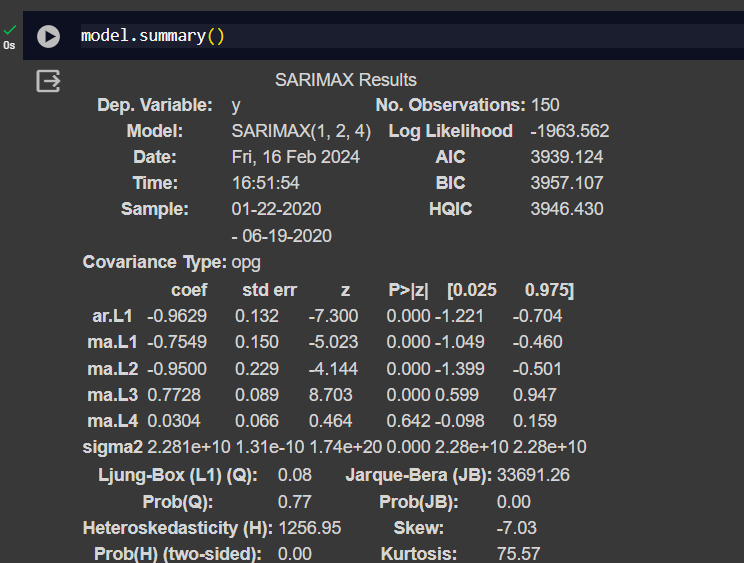
Jupyter is extensible and modular: write plugins that add new components and integrate with existing ones.

* + **Spyder**: Spyder, the Scientific Python Development Environment, is a free integrated development environment (IDE) and open-source scientific environment that is included with Anaconda written in Python, for Python, and designed by and for scientists, engineers, and data analysts.

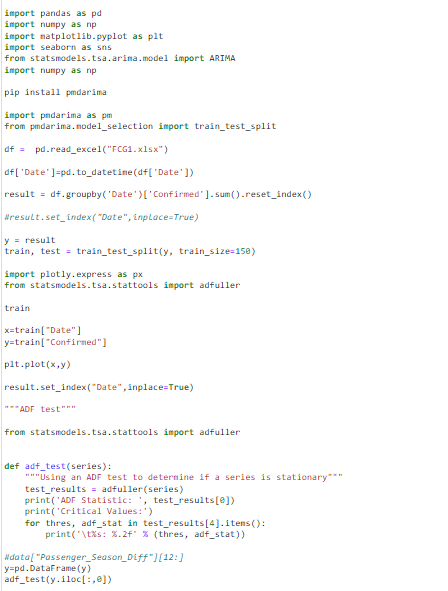
**pmdarima Model:**



## Model Summary –



## Seasonality Test-



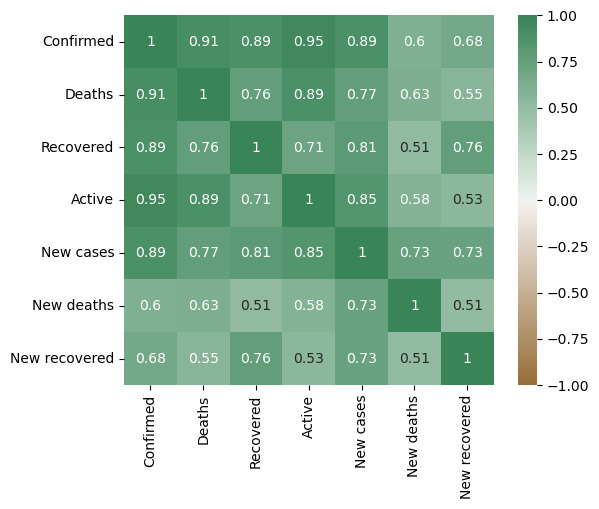
**Chapter 5**

**Results**

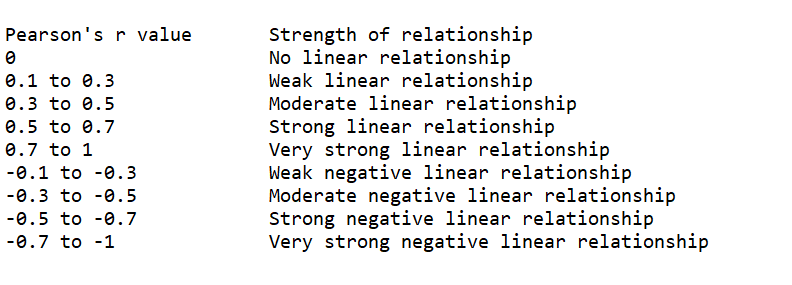
## Correlation Matrix Code



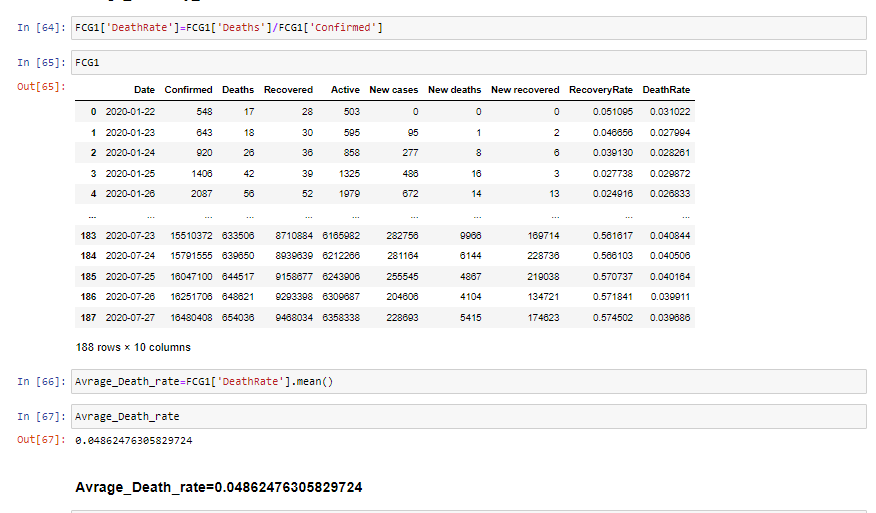
## Correlation Matrix HeatMap



## Correlation Matrix Rule



## Average Death Rate –



**Average Recovery Rate-**

**A screenshot of a computer

Description automatically generated**

# Chapter 6 Conclusion

## 6.1 Conclusion

* In conclusion, utilizing PMDARIMA (Auto ARIMA) for COVID-19 data analysis and forecasting offers several advantages. PMDARIMA automates the process of identifying optimal SARIMA parameters, making it easier and more efficient to build accurate models. By leveraging PMDARIMA, analysts can quickly assess the time series data, identify patterns, and generate reliable forecasts for COVID-19 cases.
* However, it's important to recognize the limitations of PMDARIMA and SARIMA models in general. While they provide valuable insights into temporal patterns and trends, they may not fully capture the complex dynamics of the pandemic, including the impact of interventions, behavioural changes, and other external factors. Therefore, integrating PMDARIMA forecasts with other modelling techniques and domain knowledge is essential for robust analysis and decision-making.
* In summary, PMDARIMA is a valuable tool for COVID-19 data analysis and forecasting, offering automation and efficiency in model selection and parameter tuning. When used in conjunction with other methods and a comprehensive understanding of the pandemic context, PMDARIMA can contribute to more accurate and informed predictions, aiding in public health planning and response efforts **The model can be adapted for scalability, better convergence, and better accuracy.**

## 6.2 Future Enhancement –

# Automated Model Selection: pmdarima provides functionality for automated model selection, which can streamline the process of identifying the best SARIMA parameters for the data. Future enhancements could focus on improving the efficiency and accuracy of this automated selection process, potentially incorporating advanced algorithms or heuristics.

# Model Tuning: Enhancements could be made to enable more advanced model tuning options within pmdarima, allowing users to fine-tune model parameters and optimize model performance for specific COVID-19 datasets.

# Integration with External Data Sources: Integrating pmdarima with external data sources, such as demographic data, mobility trends, or healthcare capacity, could enhance the accuracy of COVID-19 forecasts by incorporating additional relevant information into the modeling process.

# Visualization Tools: Enhancements to visualization capabilities within pmdarima could help users better understand the results of their analysis and forecasts, enabling more insightful interpretation of the data.

# Real-Time Forecasting: Future enhancements could focus on enabling real-time forecasting capabilities within pmdarima, allowing users to generate up-to-date forecasts as new COVID-19 data becomes available.

# Model Interpretability\*\*: Enhancements could be made to improve the interpretability of SARIMA models generated by pmdarima, providing users with insights into the underlying patterns and dynamics driving the COVID-19 data.

# Scalability\*\*: Improvements to the scalability of pmdarima could enable the analysis and forecasting of COVID-19 data at larger scales, accommodating datasets from regions with larger populations or more complex dynamics.

# Chapter 7 References

1. Arora, S., & Kumar, S. (2020). Modeling the dynamics of COVID-19 pandemic in India. MedRxiv. https://doi.org/10.1101/2020.04.10.20060413
2. Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., & Lessler, J. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. Annals of Internal Medicine, 172(9), 577-582.
3. Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., ... & Xing, X. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. New England Journal of Medicine, 382(13), 1199-1207.
4. Zhang, S., Diao, M., Yu, W., Pei, L., Lin, Z., & Chen, D. (2020). Estimation of the reproductive number of novel coronavirus (COVID-19) and the probable outbreak size on the Diamond Princess cruise ship: A data-driven analysis. International Journal of Infectious Diseases, 93, 201-204.
5. Prem, K., Liu, Y., Russell, T. W., Kucharski, A. J., Eggo, R. M., Davies, N., ... & Jit, M. (2020). The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: A modelling study. The Lancet Public Health, 5(5), e261-e270.
6. Biggerstaff, M., & Cauchemez, S. (2020). Real-time estimation of the reproduction number of COVID-19 in any country. medRxiv. https://doi.org/10.1101/2020.01.23.20018549
7. Kucharski, A. J., Russell, T. W., Diamond, C., Liu, Y., Edmunds, J., Funk, S., ... & Munday, J. D. (2020). Early dynamics of transmission and control of COVID-19: a mathematical modelling study. The Lancet Infectious Diseases, 20(5), 553-558.
8. Ferguson, N., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., ... & Ghani, A. (2020). Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. Imperial College London.
9. Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Mellan, T. A., Coupland, H., ... & Bhatt, S. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature, 584(7820), 257-261.
10. 10. Roosa, K., Lee, Y., Luo, R., Kirpich, A., Rothenberg, R., Hyman, J. M., ... & Chowell, G. (2020). Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020. Infectious Disease Modelling, 5, 256-263. https://www.kaggle.com/code/zabihullah18/impact-of-covid-19
11. https://github.com/topics/covid-19-data-analysis
12. https://github.com/topics/covid-19-prediction
13. <https://github.com/praveenb7/COVID-19-Data-Analysis.git>
14. <https://github.com/KenanStredic/Data-Analysis-Project.git>
15. <https://github.com/navjotsinghprince/Covid-19.git>
16. <https://github.com/ClaudiaRaffaelli/COVID-19-analysis.git>