Project Report

On

AIR QUALITY ANALYSIS

Submitted

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in

APPLIED STATISTICS AND DATA ANALYTICS

by

VINITHA V

(Register No. SM21AS021)

(2021-2023)

Under the Supervision of

ANAKHA KURIAKOSE

DEPARTMENT OF MATHEMATICS AND STATISTICS

ST. TERESA'S COLLEGE (AUTONOMOUS)

ERNAKULAM, KOCHI - 682011

APRIL 2023

ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM

CERTIFICATE

This is to certify that the dissertation entitled, **AIR QUALITY ANALYSIS** is a bonafide record of the work done by Ms. **VINITHA V** under my guidance as partial fulfillment of the award of the degree of **Master of Science in Applied Statistics and Data Analytics** at St. Teresa's College (Autonomous), Ernakulam affiliated to Mahatma Gandhi University, Kottayam. No part of this work has been submitted for any other degree elsewhere.

Date:
Place: Ernakulam

ANAKHA KURIAKOSE
Assistant Professor,
Department of Mathematics and Statistics,
St. Teresa's College(Autonomous),
Ernakulam.

Smt. Betty Joseph
Associate Professor,
Department of Mathematics and Statistics,
St. Teresa's College(Autonomous),
Ernakulam.

External Examiners

1:.............................. 2: ....

DECLARATION

I hereby declare that the work presented in this project is based on the original work done by me under the guidance of **ANAKHA KURIAKOSE**, Assistant Professor, Department of Mathematics and Statistics, St.Teresa's College(Autonomous), Ernakulam and has not been included in any other project submitted previously for the award of any degree.

Ernakulam.	**VINITHA V**
Date:	**SM21AS021**

ACKNOWLEDGEMENTS

I take this oppurtunity to thank everyone who has encouraged and supported me to carry out this project.

I am very grateful to my project guide Ms.Anakha Kuriakose for her immense help during the period of work.

In addition, I acknowledge with thanks to the Department for all the valuable support and guidance that has significantly contributed to the successful completion of this project.

I would also like to thank the HoD for her valuable suggestions and critical examination of the project.

Ernakulam. VINITHA V
Date: **SM21AS021**

ABSTRACT

With the rapid growth of economy and frequently occurring instances of air pollution, air quality has become a serious issue of discussion among the public. Air quality evaluation and air pollution control is becoming increasingly important. Air Quality Index (AQI) is an essential measure of air pollution evaluation, which describes the degree of air pollution and its effect on human health.

In this project, we perform the classification of 27 air quality monitoring stations across the state of Kerala using hierarchical and non-hierarchical clustering methods based on the ‘Ambient Air Quality Monitoring data’ for the year 2020. This project also proposes various AQI prediction models based on the AQI dataset of the city of Thiruvananthapuram for the year 2020 and compares the accuracy of these models based on RMSE scores.Comparison of the RMSE scores indicate that the Random Forest model performs better than the other models.

Contents

CERTIFICATE ii

DECLARATION iii

ACKNOWLEDGEMENTS iv

ABSTRACT v

CONTENTS vi

1 INTRODUCTION **1**

1.1 Objectives 2

2 LITERATURE REVIEW **3**

3 CLUSTER ANALYSIS **6**

3.1 Definition 6

3.2 Hierarchical Clustering 6

3.2.1 Dendrogram 7

3.3 Non-Hierarchical Clustering 7

3.3.1 K-Means Clustering 8

4 TIME SERIES ANALYSIS **9**

4.1 Time Series - Definition 9

4.2 Stationary Time Series 9

4.3 Non-Stationary Time Series 10

4.4 Autocorrelation Function (ACF) 10

4.5 Partial Autocorrelation Function (PACF) 10

4.6 Autoregressive Integrated Moving Average (ARIMA) Model 11

4.7 Augmented Dickey - Fuller Test 12

4.8 Machine Learning - Definition 12

4.9 Random Forest Regressor 13

4.10 Linear Regression 14

4.11 Long Short Term Memory Networks (LSTM) 14

5 DATA DESCRIPTION AND DATA ANALYSIS 15

5.1 Dataset 15

5.2 Data Preprocessing and Descriptive Analysis 17

5.3 Cluster Analysis 18

5.3.1 Hierarchical Clustering 18

5.3.2 K-Means Clustering 20

5.3.3 Result 21

5.4 Time Series Analysis using ARIMA Model 21

5.4.1 Time series plot of the AQI data 21

5.4.2 Decomposition of Time series 21

5.4.3 ACF and PACF 22

5.4.4 Augmented Dickey-Fuller Test 23

5.4.5 Model Building and Prediction 24

5.4.6 Result 26

5.5 Time Series Forecasting using Machine Learning Models 26

5.5.1 Random Forest Regressor Model 28

5.5.2 Linear Regression Model 29

5.5.3 Result 29

5.6 Time Series Forecasting using LSTM 30

5.6.1 Analysis 30

5.6.2 Result 32

6 CONCLUSION **33**

REFERENCES 34

Chapter 1

INTRODUCTION

Air pollution has become one of the foremost and grave public health and environmental anxiety in most evolving countries. Maintaining good air quality is important for protecting public health, preserving natural resources, and ensuring sustainable economic development. Air pollution can also have adverse effects on ecosystems, wildlife and plantlife, and contribute to climate change, acid rain, and ozone depletion. Air quality analysis and prediction is becoming increasingly important as air pollution continues to be a major global health and environmental problem. By accurately predicting and forecasting air pollution levels, we can take proactive steps to protect human health and environment. The aim of this project is to classify 27 air quality monitoring stations across the state of Kerala based on the annual average concentrations of pollutants such as Nitrogen dioxide (NO2), Sulphur dioxide (SO2), and Particulate Matter (PM10), during the year 2020. This project also considers the Air Quality Index (AQI) values of Thiruvananthapuram for the year 2020 and makes an attempt to

propose different models for the prediction of AQI.

1.1 Objectives

- **To find grouping patterns among air quality monitoring stations in Kerala using Cluster Analysis.**
- **To analyse and predict the Air Quality Index values using various Time series forecasting models.**
- **To compare the accuracy of the forecasting models.**

Chapter 2

LITERATURE REVIEW

David Núñez-Alonso, Luis Vicente Pérez-Arribas, Sadia Manzoor and Jorge O. Cáceres. Statistical Tools for Air Pollution Assessment: Multivariate and Spatial Analysis Studies in the Madrid Region reports the distribution of pollutants in the Madrid city and province from 22 monitoring stations during 2010 to 2017. Statistical tools were used to interpret and model air pollution data. [1]

Xing Wang, Zilin Wang, Min Guo, Wei Chen and Huan Zhang. Research on Air Quality Evaluation based on Principal Component Analysis evaluates the air quality of a large city in Beijing- Tianjin Hebei Area in recent 10 years and identifies influencing factors. [2]

Humaib Nasir, Kirti Goyal, Dolonchapa Prabhakar. Review of Air Quality Monitoring: Case Study of India provides insight details about current situation of air quality across various cities present in India, alongwith countless origins and effects of air pollution. [3]

B Nitesh Vamshi, Divyanshu Mishra, Punya Asthana, Namit Khanduja. Air Quality Analysis- Analysis of India's Air Quality using Anal-

ysis tools examines the changes of the concentration of air pollutants based on data which has been made publicly available by CPCB. Descriptive analysis has been used to study various trends of air pollutants at hourly and daily levels of various stations across multiple cities in India. [4]

Shubhankar Rawat. India Air Quality Data Analysis analyses air quality data of India using python. [5]

Sudesh Chaudhary, Sishil Kumar, Rimpi Antil, Sudesh Yadav. Air Quality Before and After COVID-19 Lockdown Phases Around New Delhi, India tries to assess and understand the impact of four different lockdown phases on five air pollutants compared to before lockdown at 13 air monitoring stations in and around New Delhi. [6]

Disha Sharma, Denise Mauzerall. Analysis of Air Pollution Data in India between 2015 and 2019 presents the comprehensive analysis of government air quality observations from 2015–2019 for PM10, PM2.5, SO2, NO2 and O3 from the Central Pollution Control Board (CPCB) Continuous Ambient Air Quality Monitoring (CAAQM) network and the manual National Air Quality Monitoring Program (NAMP), as well as PM2.5 from the US Air-Now network. Also addresses the inconsistencies and data gaps in datasets using a rigorous procedure to ensure data representativeness. [7]

Yue-Shan Chang, Hsin-Ta Chaio, Satheesh Abimannan, Yo-Ping Huang, Yi-Ting Tsai,Kuan-Ming Lin. An LSTM aggregated model for air pollution forecasting proposes an Aggregated Long Short- Term Memory (ALSTM) based on the LSTM deep learning method. The results reveal

that the proposed aggregate model can effectively improve accuracy of prediction. [8]

A.Gnana Soundari, J.Gnana Jeslin.M.E, Akshaya.A.C. Indian Air Quality Prediction and Analysis using Machine Learning proposes a model to predict the air quality index of a given area based on historical data of previous years and predicting over a particular year as a Gradient decent boosted multivariate regression problem. [9]

Yun-Chia Liang, Yona Maimury, Angela Chen, Josue Rodolfo Cuevas Juarez. Machine Learning-Based Prediction of Air Quality studies a series of experiments using datasets for three different regions to obtain the best prediction performance from stacking ensemble, AdaBoost and random forest models. [10]

Chapter 3

CLUSTER ANALYSIS

3.1 Definition

Cluster analysis classifies a set of observations into two or more mutually exclusive unknown groups based on combinations of interval variable. It seeks to identify a set of groups which both minimize variation within the group and maximize variation between groups. The aim is to establish a set of groups or clusters such that the elements within the cluster are more similar to each other compared to the elements in other clusters. There are mainly two types of clustering methods:

- Hierarchical
- Non-Hierarchical

3.2 Hierarchical Clustering

A hierarchical clustering algorithm works on the principle of grouping data objects into a hierarchy or "tree" of clusters. Hierarchical clustering methods proceed by a series of successive merges or a series

of successive divisions. There are two types of hierarchical clustering methods:

- **Agglomerative Clustering or Agglomerative Nesting (AGNES)**
- **Divisive Clustering or Divisive Analysis (DIANA)**

The Agglomerative Clustering algorithm starts by treating each object as a singleton cluster. Then, pairs of clusters are successively merged until all clusters have been merged into a single big cluster containing all data objects. Divisive Clustering works in the reverse manner with all data objects contained in a single big cluster and gradually being separated into groups of clusters until each object is in a singleton cluster.

3.2.1 Dendrogram

A dendrogram is a tree-like structure used for visualization of clustering partitions and is an important result of hierarchical clustering. It lists all samples and indicates at what level of similarity any two clusters were joined. Dendrogram is also a useful tool for determining the cluster number. A sudden increase in the difference between adjacent steps will indicate the appropriate number of clusters to be considered.

3.3 Non-Hierarchical Clustering

Non-hierarchical cluster analysis forms a grouping of a set of units into a pre-determined number of groups, using an iterative algorithm that optimizes a chosen criterion starting from an initial classification, units are transferred from one group to another or swapped with units from

other groups until no further improvement can be made to the criterion value. [11] One of the most popular non-hierarchical clustering method is K-means method.

3.3.1 K-Means Clustering

K-means is an example of non-hierarchical clustering method. The principle of k-means algorithm is to assign each of the 'n' data points to one of the 'k' clusters, where 'k' is a user-defined parameter. The objective is to maximize the homogeneity within the clusters and also to maximize the differences between the clusters. The homogeneity and differences are measured in terms of the distance between the objects or points in the data set. [12]. An important step in K-means clustering is to select the appropriate number of clusters. This can be determined using Elbow method. The location of a bend in the elbow plot is generally considered as an indicator of the appropriate number of clusters.

Chapter 4

TIME SERIES ANALYSIS

4.1 Time Series - Definition

A time series is a set of observations Z_t, each one being recorded at a specific time t. The main objective of time series analysis is forecasting, i.e., predicting future values. Good forecasting helps analysts to take necessary actions so as to control a given process. A given time series can be

- **Stationary**
- **Non-Stationary**

4.2 Stationary Time Series

A time series X_t is said to be stationary if the joint distribution of (X_{t_1} , X_{t_2} , ... , X_{t_n}) is identical to that of ($X_{t_{1+h}}$, $X_{t_{2+h}}$, ... , $X_{t_{n+h}}$) for all (t_1,t_2, . . . , t_n) and h, where n is an arbitrary positive integer and (t_1,t_2, . . . , t_n) is a collection of n positive integers.

4.3 Non-Stationary Time Series

A time series that fails to be stationary is called a non-stationary time series. There are several ways to transform a non-stationary data into a stationary time series data. These include method of moving averages, method of least square estimation and method of differencing. In the present study, method of differencing is applied to non-stationary data to transform it into stationary data.

4.4 Autocorrelation Function (ACF)

Autocorrelation is a mathematical representation of the degree of similarity between a given time series and a lagged version of itself over successive time intervals. [13] The autocorrelation coefficient ρ_k at lag k measures the correlation between two values X_k and X_{t-k} a distance k apart. The covariance between X_t and X_{t-k} is known as the autocovariance function at lag k. The correlation coefficient between X_t and X_{t-k} is called Autocorrelation Function (ACF) at lag k, and is given by,

$$\rho_k = Corr(X_t, X_{t-k}) = \frac{Cov(X_t, X_t - k)}{\sqrt{V(X_t) * V(X_{t-k})}} \tag{4.1}$$

4.5 Partial Autocorrelation Function (PACF)

PACF measures the extent of relationship between the current values of the variable with the earlier values while holding the effect of all other time lags constant. PACF can be used to identify the order 'p' of an autoregressive (AR) process. The PACF of an AR(p) process cuts off

at lag p so that the correct order is assessed as that value of p beyond which the sample values of partial autocorrelations are not significantly different from zero. [14] The Partial Autocorrelation Function (PACF) of a given time series X_t is the partial correlation coefficient between X_t and X_{t+h} obtained by fixing the effects of X_{t+1} , X_{t+2} , ... , X_{t+h-1}

4.6 Autoregressive Integrated Moving Average (ARIMA) Model

The ARIMA model is a useful statistical method of time series analysis defined for stationary time series for the analysis of longitudinal data with a correlation among neighbouring observations. ARIMA model can be split into smaller components, namely:

- AR : Autoregressive model which represents a type of random process. The output of the model is linearly dependent on its own previous value.
- MA : Moving Average model in which output is linearly dependent on the current and various past observations of the stochastic term.
- I : Integrated , here means the differencing step to generate stationary time series data, i.e., removing the trend and seasonal component.

ARIMA model is generally denoted as ARIMA(p,d,q) where p,d and q are the lag order of AR(p), degree of differencing and the order of MA(q) process respectively.

A process X_t is said to be an ARIMA(p,d,q) process if

$$Y_t = (1 - B)^d X_t \tag{4.2}$$

is a stationary ARMA(p,q) process ; d is a non-negative integer.

i.e, A process X_t is said to be an ARIMA(p,d,q) process if

$$\nabla^d X_t = (1 - B)^d X_t \tag{4.3}$$

is ARMA(p,q)

4.7 Augmented Dickey - Fuller Test

Augmented Dickey - Fuller Test is used to test the stationarity of a time series. The hypotheses to be tested is given by

H_0 : The series is non-stationary

v/s

H_1 : The series is stationary

The test statistic is given by,

$$DE_t = \frac{\gamma}{SE(\gamma)} \tag{4.4}$$

4.8 Machine Learning - Definition

A computer program is said to learn from experience E with respect to some class of tasks T and performance P, if its performance at tasks T as measured by P improves with experience E.

The basic machine learning process can be divided into three parts:

- Data Input: Past data is utilized as a basis for future decision making.
- Abstraction: Input data is represented in a broader way through underlying algorithm.
- Generalization: Abstracted representation is generalized to form a framework for making decisions.

Machine learning can be classified into three broad categories :

- Supervised Learning: A machine predicts the class of unknown objects based on prior class-related information on similar objects.
- Unsupervised Learning: A machine finds patterns in unknown objects by grouping similar objects together.
- Reinforcement Learning: A machine learns to act on its own to achieve the given goals. [15].

4.9 Random Forest Regressor

Random forest regressor is a type of ensemble learning algorithm used for regression tasks. It is a combination of multiple decision trees, where each tree is trained on a randomly selected subset of training data and a random subset of features. The final prediction is then made by aggregating the predictions of all the trees in the forest. Random forest regressor is a powerful algorithm which is particularly useful when dealing with large datasets and when the relationships between the features and the target variable are non-linear. It is also robust to outliers and can handle missing values in the data.

A random forest model can be used to perform time series forecasting by converting the time series data into a supervised learning problem.

4.10 Linear Regression

In linear regression, the objective is to predict numerical features. The underlying predictor variable and the target variable are continuous in nature. In case of linear regression, a straight line relationship is 'fitted' between the predictor variable and target variable, using the statistical concept of least squares method. As in the case of least square method, the sum of square of error between actual and predicted values of the target variable is tried to be minimized. Linear regression can be used to analyse time series data by treating time as a predictor variable. [15].

4.11 Long Short Term Memory Networks (LSTM)

LSTM is an artificial recurrent neural network (RNN) used in deep learning and can process entire sequences of data. Due to the model's ability to learn long term sequences of observations, LSTM has become a trending approach to time series forecasting. Using LSTM, time series forecasting models can predict future values based on previous, sequential data.This provides greater accuracy which results in better decision making. [16].

Chapter 5

DATA DESCRIPTION AND DATA ANALYSIS

5.1 Dataset

Two datasets have been considered for this project, both of which have been taken from https://cpcb.nic.in

The first dataset is the National Ambient Air Quality Monitoring Data for the year 2020. [NAMP data]. It contains 27 rows and 5 columns/attributes given by,

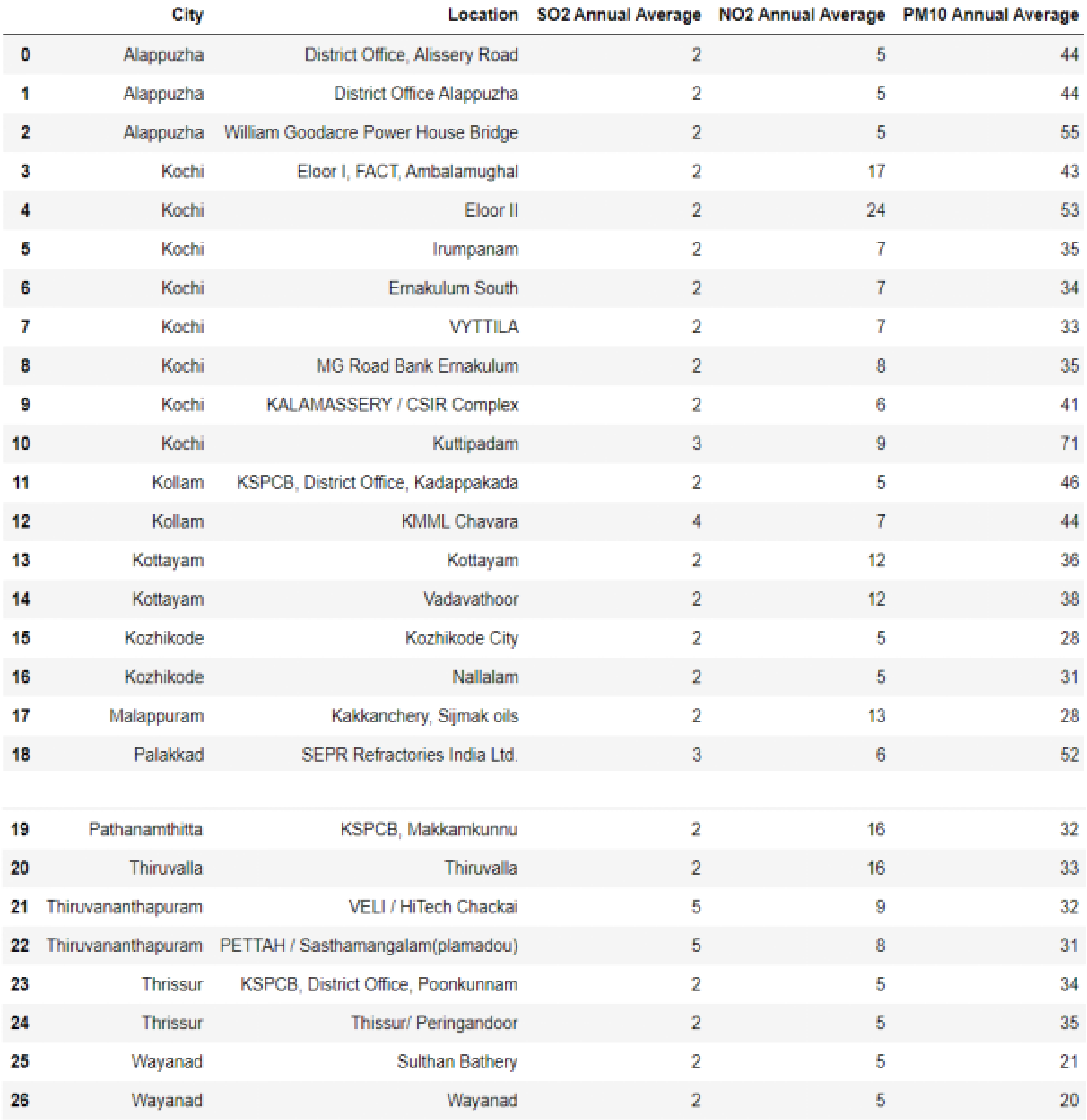
i. City : City where the monitoring stations are located

ii. Location: Location of the monitoring station

iii. SO_2 : Annual average concentration of SO_2

iv. NO_2 : Annual average concentration of NO_2

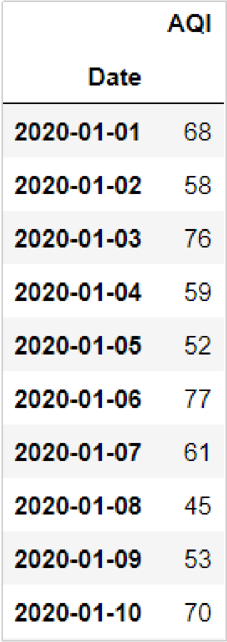
v. PM_{10} : Annual average concentration of PM_{10}



	City	Location	SO2 Annual Average	NO2 Annual Average	PM10 Annual Average
0	Alappuzha	District Office, Alissery Road	2	5	44
1	Alappuzha	District Office Alappuzha	2	5	44
2	Alappuzha	William Goodacre Power House Bridge	2	5	55
3	Kochi	Eloor I, FACT, Ambalamughal	2	17	43
4	Kochi	Eloor II	2	24	53
5	Kochi	Irumpanam	2	7	35
6	Kochi	Ernakulum South	2	7	34
7	Kochi	VYTTILA	2	7	33
8	Kochi	MG Road Bank Ernakulum	2	8	35
9	Kochi	KALAMASSERY / CSIR Complex	2	6	41
10	Kochi	Kuttipadam	3	9	71
11	Kollam	KSPCB, District Office, Kadappakada	2	5	46
12	Kollam	KMML Chavara	4	7	44
13	Kottayam	Kottayam	2	12	36
14	Kottayam	Vadavathoor	2	12	38
15	Kozhikode	Kozhikode City	2	5	28
16	Kozhikode	Nallalam	2	5	31
17	Malappuram	Kakkanchery, Sijmak oils	2	13	28
18	Palakkad	SEPR Refractories India Ltd.	3	6	52
19	Pathanamthitta	KSPCB, Makkamkunnu	2	16	32
20	Thiruvalla	Thiruvalla	2	16	33
21	Thiruvananthapuram	VELI / HiTech Chackai	5	9	32
22	Thiruvananthapuram	PETTAH / Sasthamangalam(plamadou)	5	8	31
23	Thrissur	KSPCB, District Office, Poonkunnam	2	5	34
24	Thrissur	Thissur/ Peringandoor	2	5	35
25	Wayanad	Sulthan Bathery	2	5	21
26	Wayanad	Wayanad	2	5	20

Figure 5.1: NAMP Data

The second dataset is the Air Quality Index (AQI) data of the city of Thiruvananthapuram from January 1st, 2020 to December 31st, 2020.



Date	AQI
2020-01-01	68
2020-01-02	58
2020-01-03	76
2020-01-04	59
2020-01-05	52
2020-01-06	77
2020-01-07	61
2020-01-08	45
2020-01-09	53
2020-01-10	70

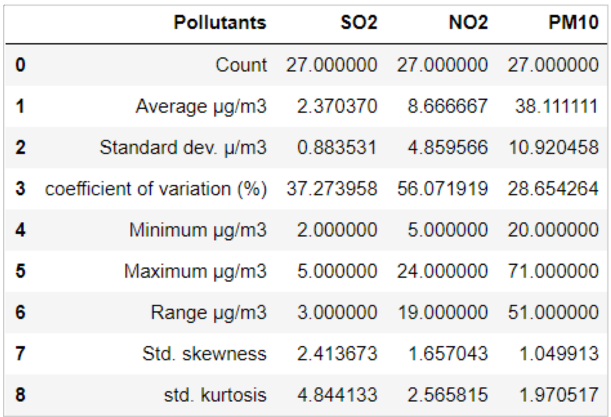
Figure 5.2: AQI Dataset

The implementation was done using Python in Jupyter Notebook.

5.2 Data Preprocessing and Descriptive Analysis

The required libraries such as pandas, NumPy, matplotlib, etc.. are imported and the datasets are loaded. Both datasets are checked for null values and neither one has any.

The descriptive analysis of the NAMP data is performed and the following results are obtained.



	Pollutants	SO2	NO2	PM10
0	Count	27.000000	27.000000	27.000000
1	Average µg/m3	2.370370	8.666667	38.111111
2	Standard dev. µ/m3	0.883531	4.859566	10.920458
3	coefficient of variation (%)	37.273958	56.071919	28.654264
4	Minimum µg/m3	2.000000	5.000000	20.000000
5	Maximum µg/m3	5.000000	24.000000	71.000000
6	Range µg/m3	3.000000	19.000000	51.000000
7	Std. skewness	2.413673	1.657043	1.049913
8	std. kurtosis	4.844133	2.565815	1.970517

Figure 5.3: Summary Statistics

Correlation Matrix

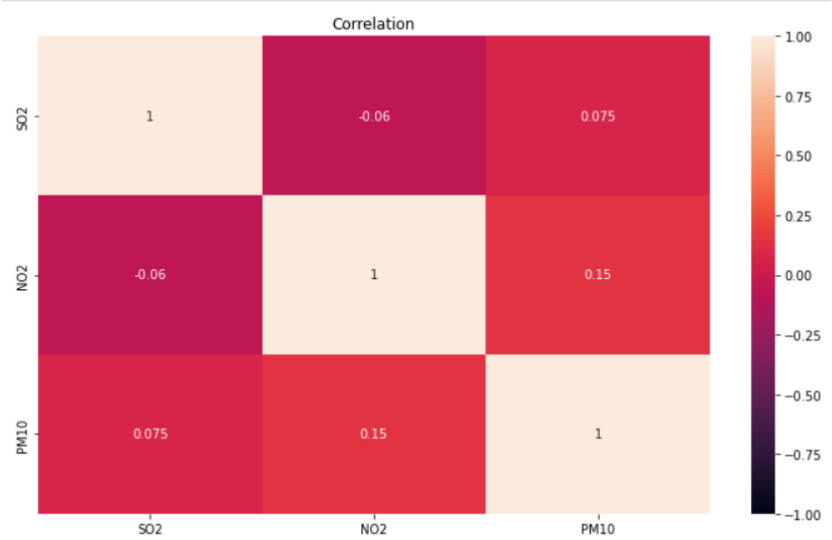


Figure 5.4: Correlation Matrix

Each cell in a correlation matrix shows the correlation between the corresponding two variables. Here we can see that the variables (pollutant concentrations) are not significantly correlated.

5.3 Cluster Analysis

5.3.1 Hierarchical Clustering

We perform hierarchical clustering on the NAMP data. The linkage used is Ward's linkage.The result can be visualized in the following dendrogram.

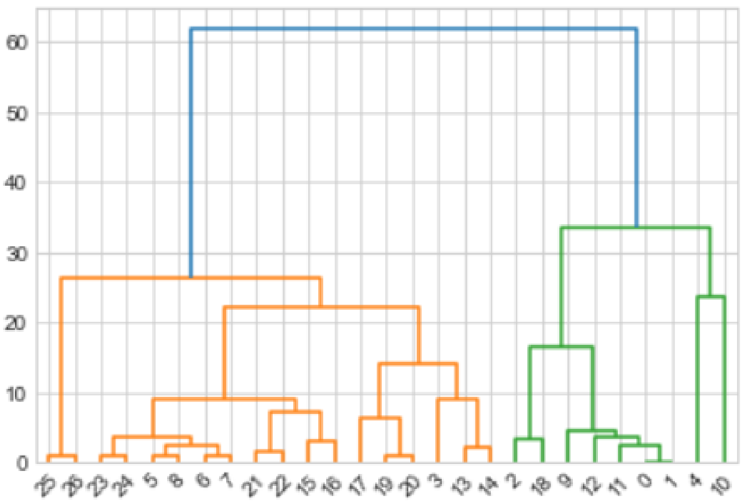


Figure 5.5: Dendrogram

From the dendrogram, it is clear that the monitoring stations are divided into two distinct clusters as shown in Table.5.1.

CLUSTER 1	CLUSTER 2
Sulthan Bathery	William Goodacre Power House Bridge
Wayanad	SEPR Refractories India Ltd.
KSPCB, District Office, Poonkunnam	KALAMASSERY/CSIR Complex
Thissur/Peringandoor	KMML Chavara
Irumpanam	KSPCB, District Office, Kadappakada
MG Road Bank Ernakulam	District Office, Alissery Road
Ernakulam South	District Office Alappuzha
VYTTILA	Eloor II
VELI/HiTech Chackai	Kuttipadam
PETTAH/Sasthamangalam(plamadou)	-
Kozhikode City	-
Nallalam	-
Kakkanchery, Sijmak oils	-
KSPCB, Makkamkunnu	-
Thiruvalla	-
Eloor I, FACT, Ambalamughal	-
Kottayam	-
Vadavathoor	-

Table 5.1: Clusters in hierarchical clustering

5.3.2 K-Means Clustering

By K-means method, we can group the 27 monitoring stations based on the 3 variables. We use Elbow method to determine the appropriate number of clusters k.

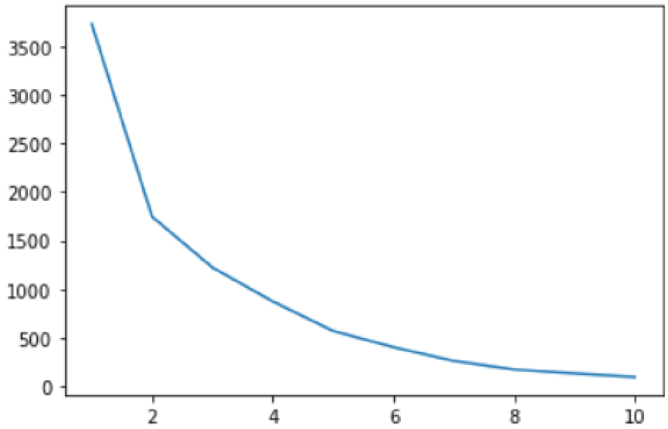


Figure 5.6: Elbow Graph

From Fig. 5.6 , we conclude that k=2

The observations are classified into clusters as shown in Table. 5.2.

CLUSTER 1	CLUSTER 2
District Office, Alissery Road	Irumpanam
District Office Alappuzha	Ernakulam South
William Goodacre Power House Bridge	VYTTILA
Eloor I, FACT, Ambalamughal	MG Road Bank Ernakulam
Eloor II	Kottayam
KALAMASSERY/CSIR Complex	Vadavathoor
Kuttipadam	Kozhikode City
KSPCB, District Office, Kadappakada	Nallalam
KMML Chavara	Kakkanchery, Sijmak oils
SEPR Refractories India Ltd.	KSPCB, Makkamkunnu
-	Thiruvalla
-	VELI/HiTech Chackai
-	PETTAH/Sasthamangalam(plamadou)
-	KSPCB, District Office, Poonkunnam
-	Thissur/Peringandoor
-	Sulthan Bathery
-	Wayanad

Table 5.2: K-Means Clusters

5.3.3 Result

Hierarchical clustering is performed and the data is classified into 2 clusters using dendrogram. In K-means clustering, appropriate number of clusters is obtained as k=2 using Elbow method. The data is then grouped into 2 clusters based on the 3 variables.

5.4 Time Series Analysis using ARIMA Model

5.4.1 Time series plot of the AQI data

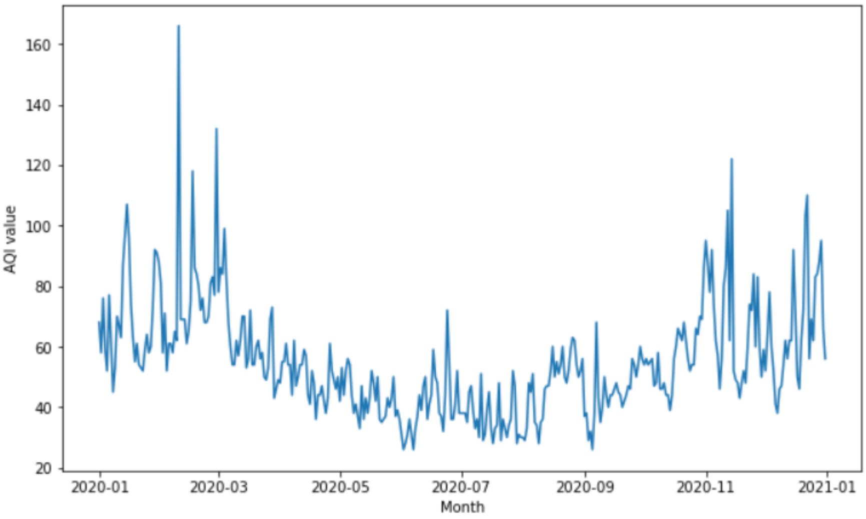


Figure 5.7: Time Series Plot

Fig.5.7. shows the time series plot for the air quality indices of Thiruvananthapuram, from January 1st,2020 to December 31st,2020.

5.4.2 Decomposition of Time series

We perform seasonal decomposition for evaluating the trend and seasonality of the time series. From Fig.5.8 it is clear that the series does not have any particular trend.

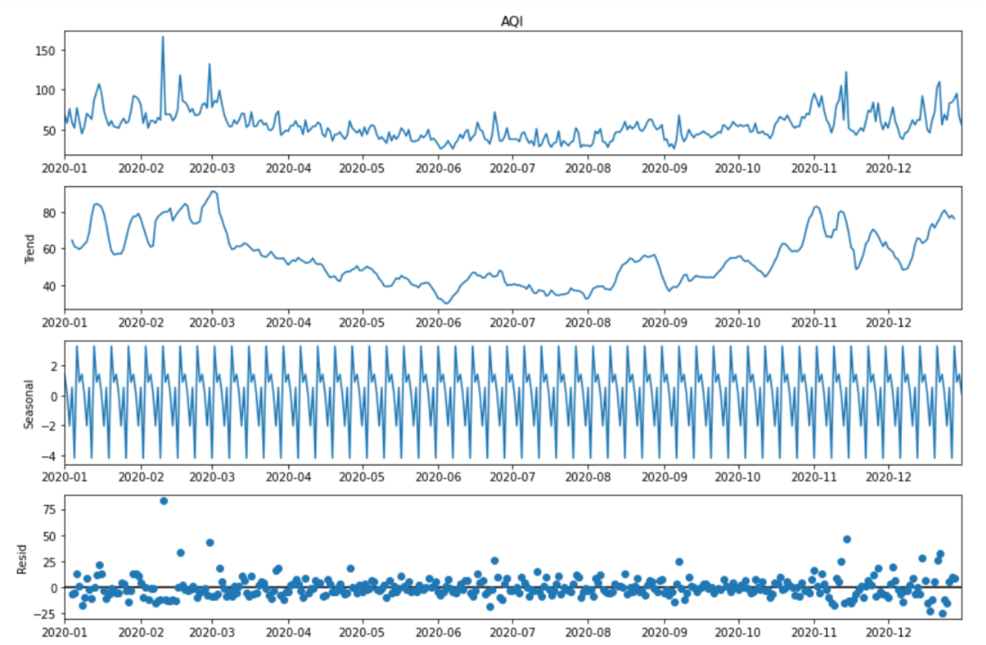


Figure 5.8: Decomposition of Time series

5.4.3 ACF and PACF

The next step is to find the ACF and PACF of the time series data. Fig.5.9 shows the ACF and PACF plots of the data.

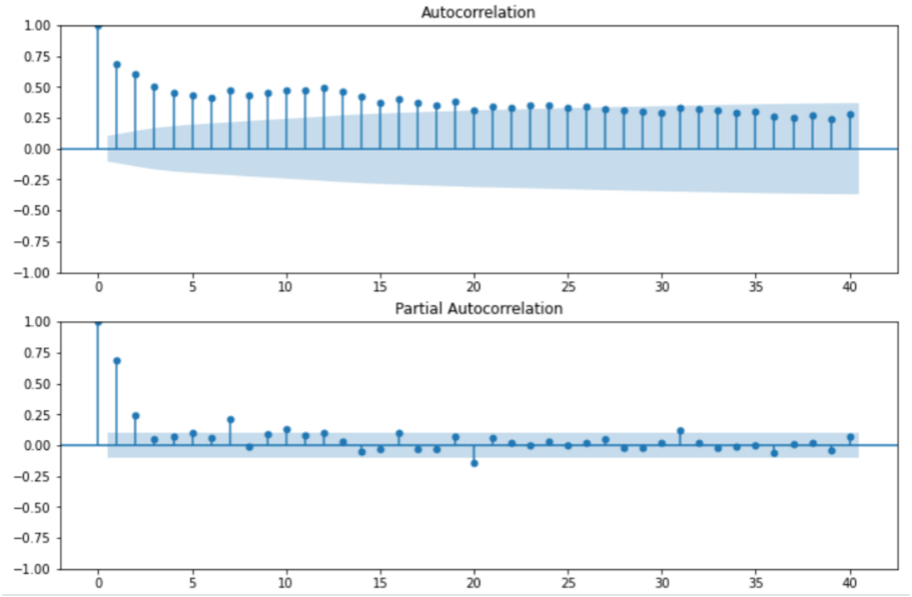
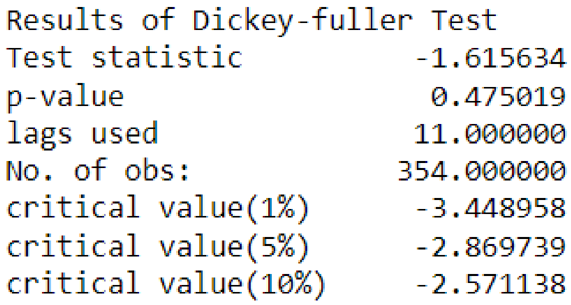


Figure 5.9: ACF and PACF plots

5.4.4 Augmented Dickey-Fuller Test

H_0 : The series is non-stationary v/s H_1 : The series is stationary. The result of the ADF test is obtained as follows.



```
Results of Dickey-fuller Test
Test statistic              -1.615634
p-value                      0.475019
lags used                   11.000000
No. of obs:                354.000000
critical value(1%)          -3.448958
critical value(5%)          -2.869739
critical value(10%)         -2.571138
```

Figure 5.10:

Since p value is greater than 0.05, we fail to reject the null hypothesis. Therefore the series is non-stationary. In order to make it stationary, the data is differenced. The time series plot after differencing is given in Fig.5.11.

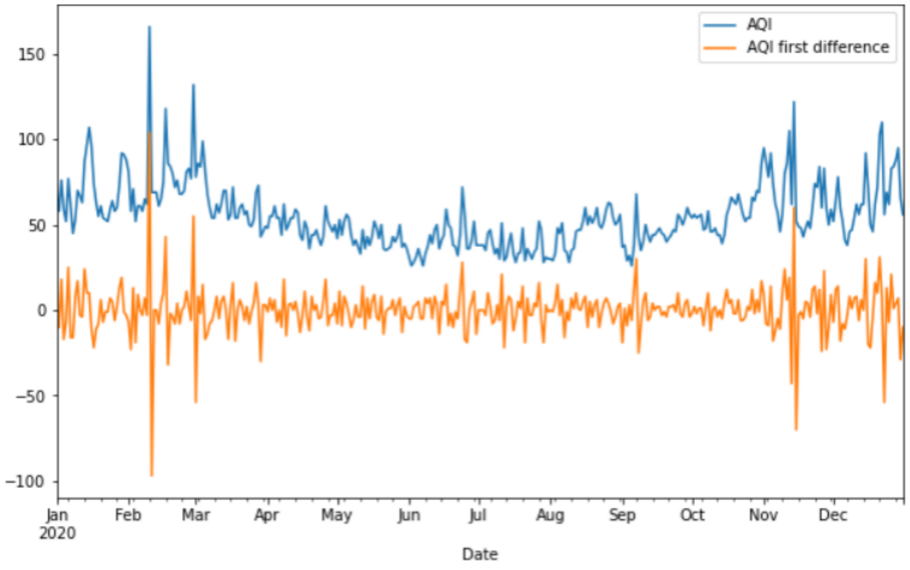
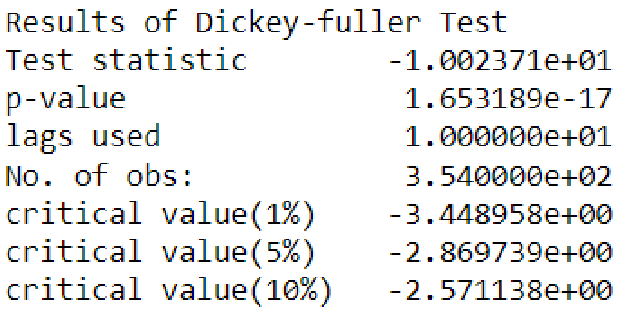


Figure 5.11: Time series plot after differencing

It is again checked for stationarity using ADF test. The following results are obtained.



```
Results of Dickey-fuller Test
Test statistic        -1.002371e+01
p-value                1.653189e-17
lags used              1.000000e+01
No. of obs:            3.540000e+02
critical value(1%)    -3.448958e+00
critical value(5%)    -2.869739e+00
critical value(10%)   -2.571138e+00
```

Figure 5.12:

Since p value is less than 0.05 we reject null hypothesis. Hence the series is stationary.

The ACF and PACF after differencing is given as shown in Fig.5.13.

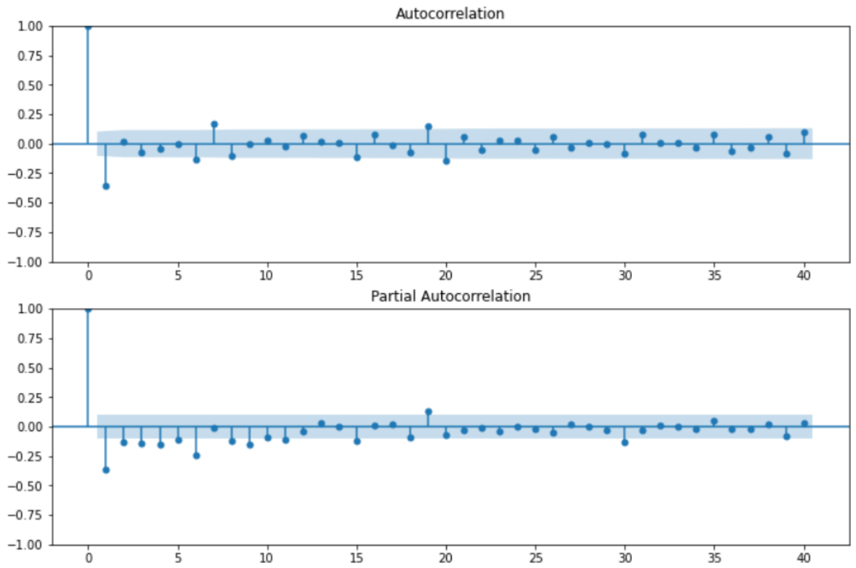
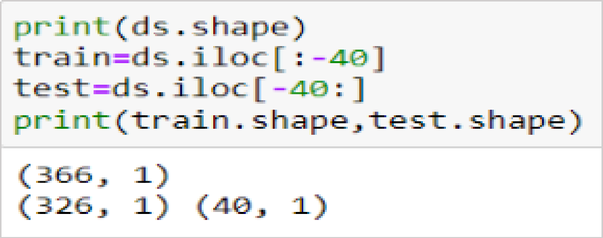


Figure 5.13: ACF and PACF plots after differencing

We can determine the Autoregressive parameter from the PACF plot and the Moving Average parameter from the ACF plot.

5.4.5 Model Building and Prediction

In order to fit the data into ARIMA(1,1,1) model, first we split the data into train and test data as shown.

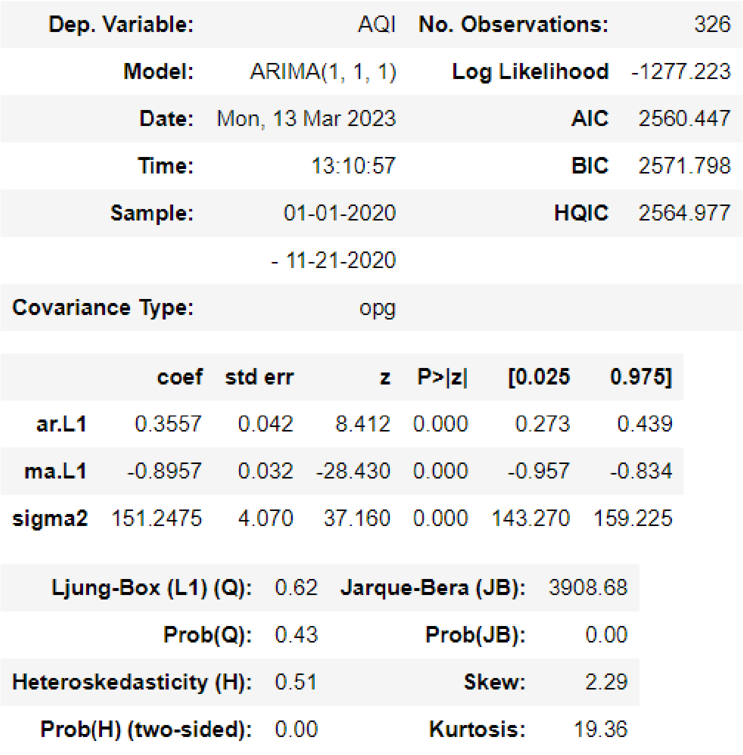


```
print(ds.shape)
train=ds.iloc[:-40]
test=ds.iloc[-40:]
print(train.shape,test.shape)
```

```
(366, 1)
(326, 1) (40, 1)
```

Figure 5.14:

Then we train the model by fitting the training data and thus obtain the following results.



Dep. Variable:	AQI	**No. Observations:**	326
Model:	ARIMA(1, 1, 1)	**Log Likelihood**	-1277.223
Date:	Mon, 13 Mar 2023	**AIC**	2560.447
Time:	13:10:57	**BIC**	2571.798
Sample:	01-01-2020	**HQIC**	2564.977
	- 11-21-2020		
Covariance Type:	opg		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
ar.L1	0.3557	0.042	8.412	0.000	0.273	0.439
ma.L1	-0.8957	0.032	-28.430	0.000	-0.957	-0.834
sigma2	151.2475	4.070	37.160	0.000	143.270	159.225

Ljung-Box (L1) (Q):	0.62	**Jarque-Bera (JB):**	3908.68
Prob(Q):	0.43	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.51	**Skew:**	2.29
Prob(H) (two-sided):	0.00	**Kurtosis:**	19.36

Figure 5.15: Fitted Model

Based on the above results, we try to predict the values of the test data. The plotted graph in Fig.5.16 draws a comparison between the predicted values and the actual values of the test data.

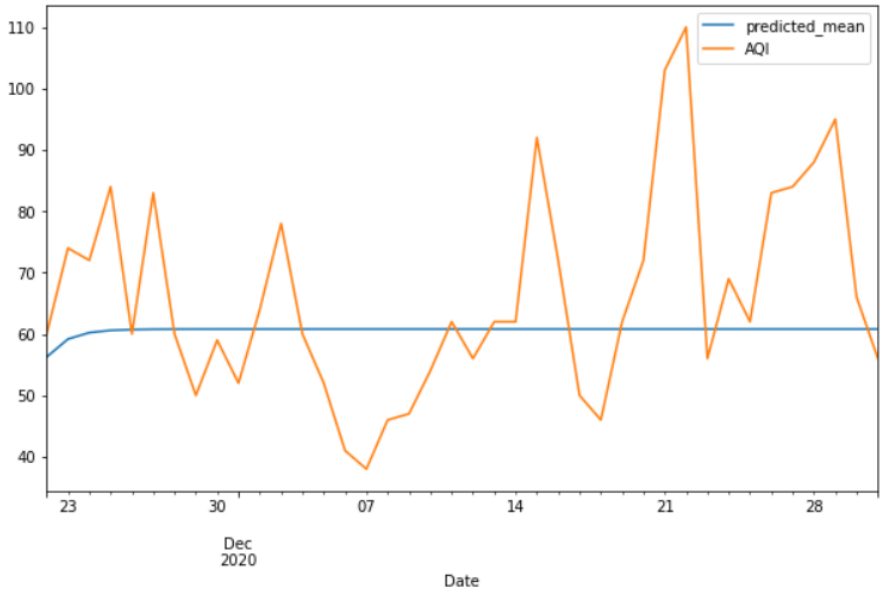


Figure 5.16: Predicted AQI Plot

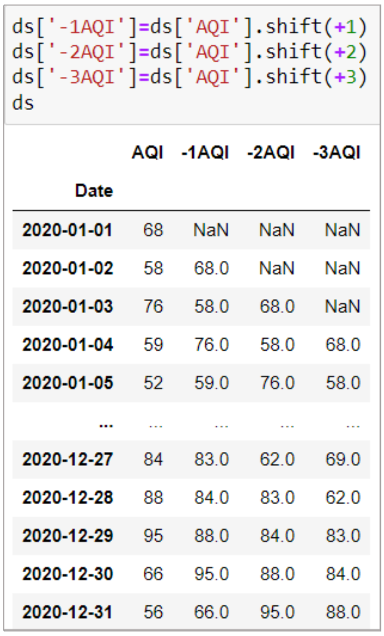
The accuracy of the model is determined using RMSE score given by, 17.490667542513705.

5.4.6 Result

A model is fitted and a comparison is drawn between the actual and predicted values.The RMSE score of the model is 17.490667542513705.

5.5 Time Series Forecasting using Machine Learning Models

In order to convert the time series data into a supervised learning problem, we shift the AQI values in the dataset by 1 unit. This gives the AQI value of the previous day based on which we can predict the AQI value of any given day. Similarly, we shift the AQI values by 2 and 3 units so as to obtain the AQI of the past 2days and 3 days based on which we can predict the AQI value for any given day.



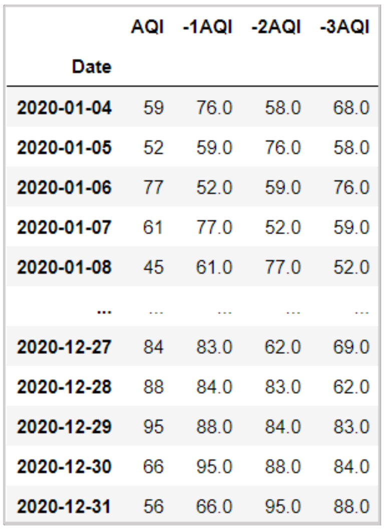
```
ds['-1AQI']=ds['AQI'].shift(+1)
ds['-2AQI']=ds['AQI'].shift(+2)
ds['-3AQI']=ds['AQI'].shift(+3)
ds
```

	AQI	-1AQI	-2AQI	-3AQI
Date				
2020-01-01	68	NaN	NaN	NaN
2020-01-02	58	68.0	NaN	NaN
2020-01-03	76	58.0	68.0	NaN
2020-01-04	59	76.0	58.0	68.0
2020-01-05	52	59.0	76.0	58.0
...	...	...	...	...
2020-12-27	84	83.0	62.0	69.0
2020-12-28	88	84.0	83.0	62.0
2020-12-29	95	88.0	84.0	83.0
2020-12-30	66	95.0	88.0	84.0
2020-12-31	56	66.0	95.0	88.0

Figure 5.17:

Fig.5.17 shows the results so obtained. Here, -1AQI = AQI of previous day, -2AQI = AQI 2 days before, -3AQI = AQI 3 days before.

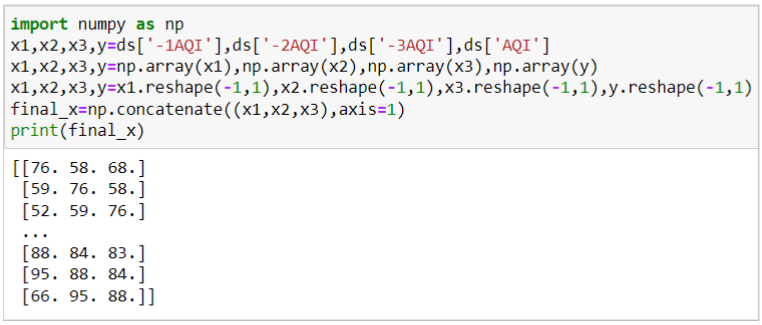
Now, Fig.5.18 shows the results after removing all the NaN values.



	AQI	-1AQI	-2AQI	-3AQI
Date				
2020-01-04	59	76.0	58.0	68.0
2020-01-05	52	59.0	76.0	58.0
2020-01-06	77	52.0	59.0	76.0
2020-01-07	61	77.0	52.0	59.0
2020-01-08	45	61.0	77.0	52.0
...	...	...	...	...
2020-12-27	84	83.0	62.0	69.0
2020-12-28	88	84.0	83.0	62.0
2020-12-29	95	88.0	84.0	83.0
2020-12-30	66	95.0	88.0	84.0
2020-12-31	56	66.0	95.0	88.0

Figure 5.18:

Next we preprocess the data before it is split into training and test data.

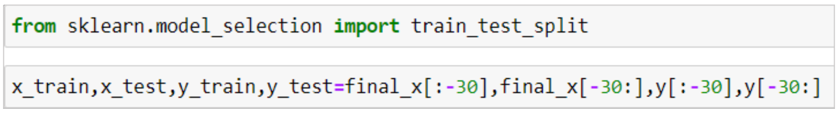


```
import numpy as np
x1,x2,x3,y=ds['-1AQI'],ds['-2AQI'],ds['-3AQI'],ds['AQI']
x1,x2,x3,y=np.array(x1),np.array(x2),np.array(x3),np.array(y)
x1,x2,x3,y=x1.reshape(-1,1),x2.reshape(-1,1),x3.reshape(-1,1),y.reshape(-1,1)
final_x=np.concatenate((x1,x2,x3),axis=1)
print(final_x)
```

```
[[76. 58. 68.]
 [59. 76. 58.]
 [52. 59. 76.]
 ...
 [88. 84. 83.]
 [95. 88. 84.]
 [66. 95. 88.]]
```

Figure 5.19:

The data is then split into training and test data as shown.



```
from sklearn.model_selection import train_test_split
```

```
x_train,x_test,y_train,y_test=final_x[:-30],final_x[-30:],y[:-30],y[-30:]
```

Figure 5.20:

5.5.1 Random Forest Regressor Model

We import random forest regressor from sklearn.ensemble in python and fit the training and test data to the model. Based on the fitted model, predictions are made and a graph is plotted comparing the actual and predicted values.

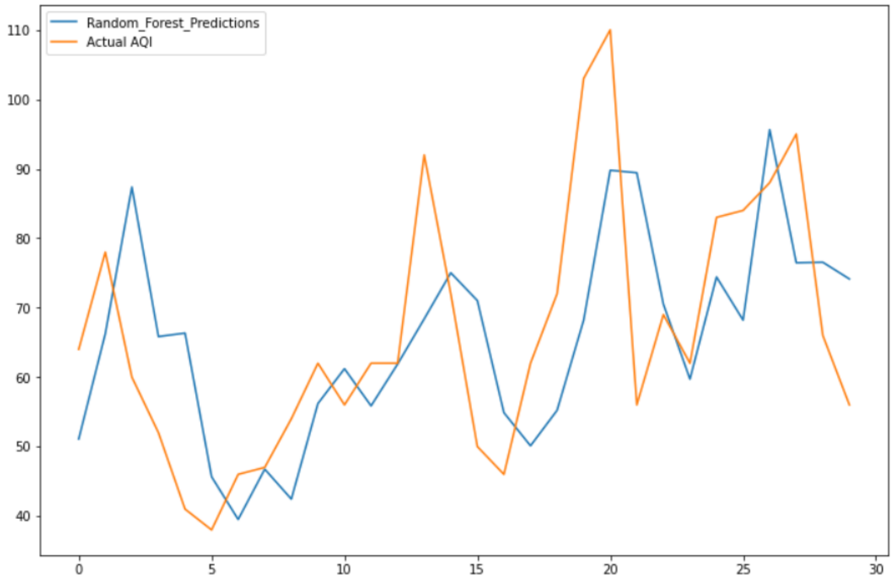


Figure 5.21: Random Forest Predictions

The model accuracy is given by the RMSE score 15.958708223040741 and the cross-validation score of the model is 8.900488715513342 .

5.5.2 Linear Regression Model

We import linear regression from sklearn.linear model in python and fit the training and test data to the model. Based on the fitted model, predictions are made and a graph is plotted comparing the actual and predicted values.

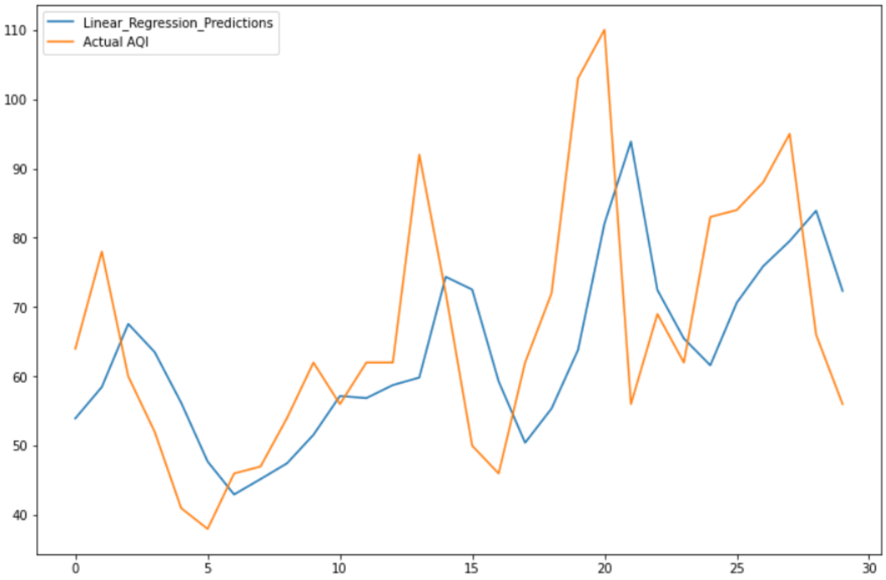


Figure 5.22: Linear Regression Predictions

The model accuracy is given by the RMSE score 17.047266468054225 and the cross-validation score of the model is 8.249761384698159

5.5.3 Result

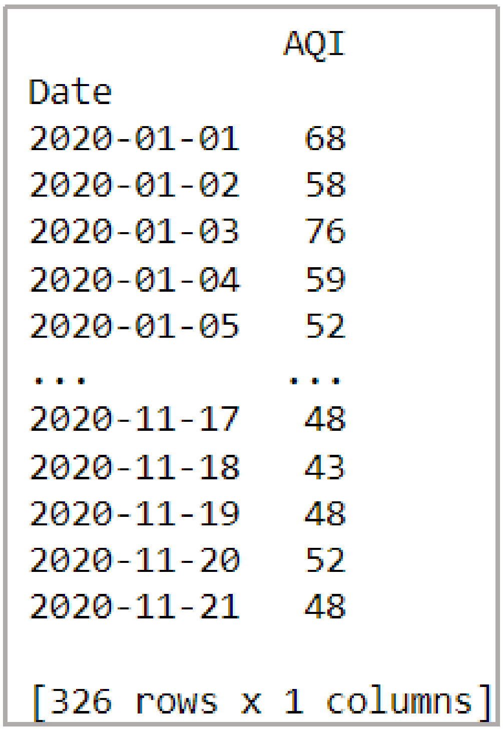
The RMSE scores and cross-validation scores of the two machine learning models- Random forest regressor and Linear regression, are determined. It is observed that the Random Forest model, with lower RMSE score, performs better than the Linear Regression model. A compari-

son of actual and predicted values of both models is drawn using plotted graphs.

5.6 Time Series Forecasting using LSTM

5.6.1 Analysis

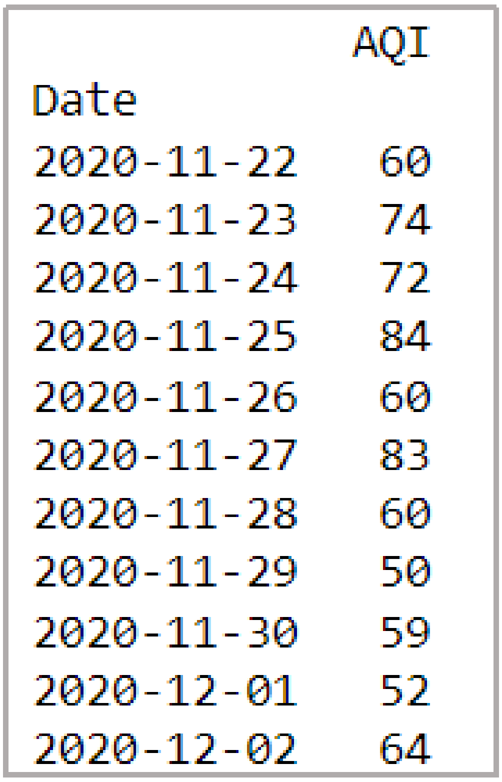
We import the necessary libraries such as pandas, numpy, tensorflow etc.. and the read the dataset in Python. The data contains 366 observations and it is split into training and testing data as shown.



```
             AQI
Date
2020-01-01    68
2020-01-02    58
2020-01-03    76
2020-01-04    59
2020-01-05    52
...          ...
2020-11-17    48
2020-11-18    43
2020-11-19    48
2020-11-20    52
2020-11-21    48

[326 rows x 1 columns]
```

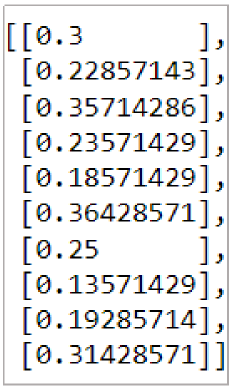
Figure 5.23: Training Data



```
             AQI
Date
2020-11-22    60
2020-11-23    74
2020-11-24    72
2020-11-25    84
2020-11-26    60
2020-11-27    83
2020-11-28    60
2020-11-29    50
2020-11-30    59
2020-12-01    52
2020-12-02    64
```

Figure 5.24: Testing Data

The next step is to pre-process the data using a MinMax Scaler to convert the dataset into a scale of 0 to 1. After that we fit the scaled data to the training set and transform both training and testing set using transform function. We obtain the values within 0 to 1 range as shown.

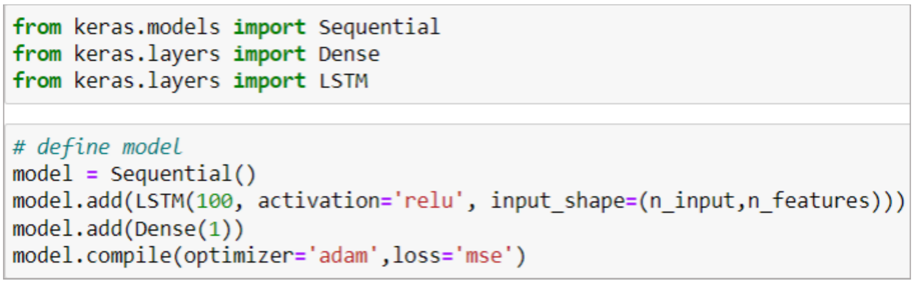


```
[[0.3       ],
 [0.22857143],
 [0.35714286],
 [0.23571429],
 [0.18571429],
 [0.36428571],
 [0.25      ],
 [0.13571429],
 [0.19285714],
 [0.31428571]]
```

Figure 5.25: Scaled Data

Next we format the data to fit the neural network model using Time series generator.

After that we define an LSTM model using the following code.



```
from keras.models import Sequential
from keras.layers import Dense
from keras.layers import LSTM
```

```
# define model
model = Sequential()
model.add(LSTM(100, activation='relu', input_shape=(n_input,n_features)))
model.add(Dense(1))
model.compile(optimizer='adam',loss='mse')
```

Figure 5.26:

Finally, we fit the model for 40 epochs. From Fig.5.27 we can see that the loss has been decreasing with every epoch.

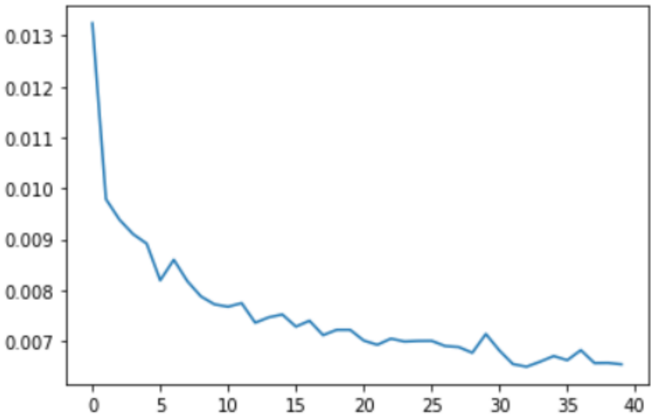
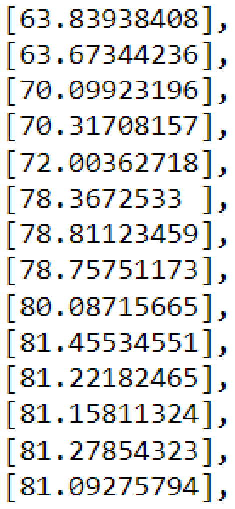


Figure 5.27: Loss per Epoch

Based on the fitted model, we make predictions for the test data.



```
[63.83938408],
[63.67344236],
[70.09923196],
[70.31708157],
[72.00362718],
[78.3672533 ],
[78.81123459],
[78.75751173],
[80.08715665],
[81.45534551],
[81.22182465],
[81.15811324],
[81.27854323],
[81.09275794],
```

Figure 5.28: Predicted Values

Fig.5.29. plots a comparison between the test values and predictions.

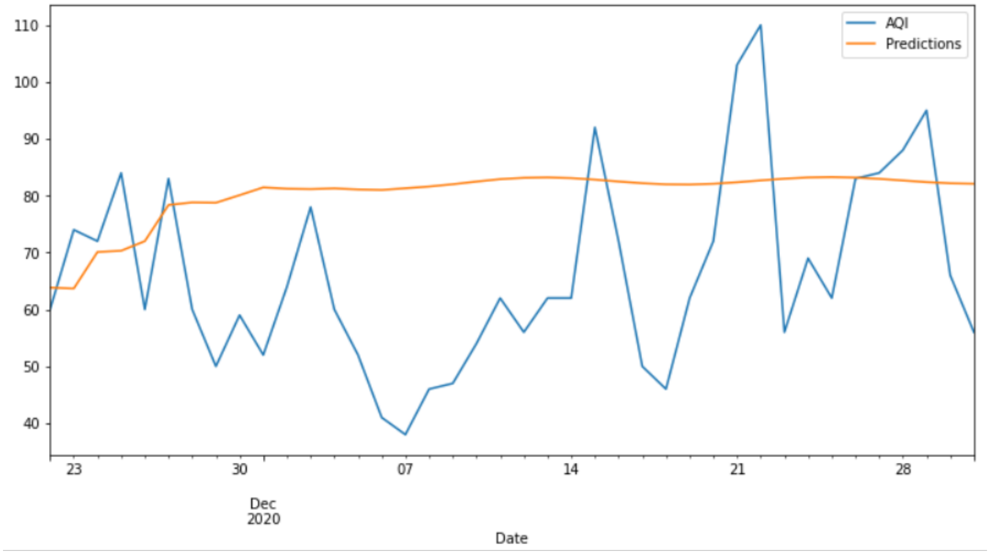


Figure 5.29: Comparison between Actual and Predicted Test Data values

The accuracy of the model is given by the RMSE score 22.437971007010812.

5.6.2 Result

A model is fitted and a comparison is drawn between the actual and predicted test values using plotted graph. The RMSE score of the model is 22.437971007010812.

Chapter 6

CONCLUSION

The objectives of the project have been achieved. The 27 air quality monitoring stations were classified into two clusters using hierarchical and non-hierarchical (k-means) clustering. Different Time series models were fitted using the Air Quality Index Data and graphs were plotted to compare actual values and predictions. Each of their respective model accuracy was determined using RMSE scores. Further comparison of the RMSE scores indicate that the Random Forest model performs better and is more accurate than the other models.

REFERENCES

[1] https://www.hindawi.com/journals/jamc/2019/9753927/

[2] https://iopscience.iop.org/article/10.1088/1755-1315/108/4/042030

[3] https://www.researchgate.net/publication/311499664_Review_of_Air_Quality_Monitoring_Case_Study_of_India

[4] https://www.researchgate.net/publication/353413170_Air_Quality_analysis

[5] https://towardsdatascience.com/india-air-pollution-data-analysis-bd7dbfe938

[6] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8276728/

[7] https://aaqr.org/articles/aaqr-21-08-oa-0204

[8] https://www.sciencedirect.com/science/article/pii/S1309104220301215

[9] https://www.ripublication.com/ijaerspl2019/ijaerv14n11spl_34.pdf

[10] https://www.mdpi.com/2076-3417/10/24/9151

[11] https://genstat22.kb.vsni.co.uk/knowledge-base/non-hierarchical-cluster-analysis/

[12] **Saikat Dutt, Subramanian Chandramouli, Amit Kumar Das "Unsupervised Learning ", Machine Learning, No.9, 247-249, Apr 2018.**

[13] https://analyticsindiamag.com/an-overview-of-autocorrelation-seasonality-an

[14] **Won Kwang Paik "Methodology for Trade Research ", United States Trade Relations with the Newly Industrializing Countries in the Pacific Basin, No.3, 47, 2019.**

[15] **Saikat Dutt, Subramanian Chandramouli, Amit Kumar Das "Introduction to Machine Learning ", Machine Learning, No.1, 6-14, Apr 2018.**

[16] `https://www.predicthq.com/events/lstm-time-series-forecasting`