TIME SERIES FORECASTING FOR DIVERSE LOCALES POLLUTION WITH APACHE SPARK

A MINI PROJECT REPORT

Submitted by

SANJJUSHRI VARSHINI R– 210420243047

PRIYADHARSHINI R- 210420243042

YOGA PRIYA K- 210420243062

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

**CHENNAI INSTITUTE OF TECHNOLOGY,
PUDUPER, KANCHEEPURAM**

ANNA UNIVERSITY: CHENNAI 600 025

DECEMBER 2022

ANNA UNIVERSITY: CHENNAI 600 025

BONAFIDE CERTIFICATE

Certified that this project report titled "**TIME SERIES FORECASTING FOR DIVERSELOCALES POLLUTION WITH APACHE SPARK**" is the bonafide work of " **SANJJUSHRI VARSHINI R (210420243047), PRIYADHARSHINI R (210420243042) AND YOGA PRIYA K (210420243062)**" who carried out the project work under my supervision.

SIGNATURE

Dr.M.GEETHA,Ph.D
HEAD OF THE DEPARTMENT

Professor
Department Of Artificial
Intelligence And Data Science.
Chennai Institute Of Technology,
Kundrathur, Chennai-600069.

SIGNATURE

Ms R.JAAZIELIAH
SUPERVISOR

Assistant Professor
Department Of Artificial
Intelligence And Data Science.
Chennai Institute Of Technology,
Kundrathur, Chennai-600069.

Submitted for the **ANNA UNIVERSITY** examination held on ____________________
at Chennai Institute of Technology, Kundrathur.

INTERNAL EXAMINER

EXTERNAL EXAMINER

ACKNOWLEDGEMENT

At this crucial moment, we thank our beloved Chairman **Shri.P.SRIRAM** and all the trust members of Chennai Institute of Technology for providing us with a plethora of facilities to successfully complete my thesis.

We would like to express our gratitude to our Principal **Dr.A.RAMESH**, who has been a pillar of spiritual power and a constant source of inspiration to us. **Dr.M.GEETHA, PhD**, Head of the Department, Chennai Institute of Technology, has provided us with invaluable advice and suggestions. We are overjoyed to express our sincere gratitude to **Dr.M.GEETHA, PhD**, our beloved project guide, for his insightful advice, outstanding encouragement and unwavering support throughout the course of our project.

We also thank the teaching and non-teaching staff members of the Information Technology Department and all our fellow students who stood with us to complete our project successfully. Last but not least we extend our deep gratitude to our beloved family members for their moral coordination, encouragement and financial support to carry out this project.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

In day-to-day life, humans face many tribulations due to unwholesome routines that induce much pollution. It is critical to understand how pollution would impact our environment and health if the same problem continues. Machine Learning for time series forecasting incorporated with Apache Spark gives more potential to work and study in our environment effectively. Apache Cassandra, a highly scalable database has been utilised for better performance. This research aids to identify what, where, and when precautions should be taken to protect human health as well as the environment. The earlier measured data of prior years is operated to compute the future circumstances of the surroundings concerning pollution in a factual method. Therefore this research paper contributes effectively, to how the future environment at various locations will be if we follow the exact as now.

Keywords: Time series forecasting, Machine Learning, Apache Spark, Apache Cassandra, Pollution

CHAPTER 1
INTRODUCTION

1.1 OVERVIEW

In the modern world, pollution must be taken seriously. As the world's population increases the pollution emitted also increases steadily. Minor Changes in the concentration of air will cause a significant hazard to people's health. Urban air tends to be more polluted than rural air because the size of pollution particles is generally large. The types of pollution discussed in this paper are Air and Noise pollution which can be further extended in future. Burning fossil fuels, which are mostly done to create electrical energy and power engines, is the major cause of air pollution. Renewable energy sources, such as solar energy, wind energy, etc., can be used in their place. When pollutants like chlorofluorocarbons (CFCs) release chlorine atoms into the atmosphere, they destroy ozone because they include chlorine atoms. Ozone molecules are significantly reduced by just one chlorine atom. The impact of Noise pollution is equal to that of air pollution. The planet is encircled by air which is our primary medium of communication affecting it will cause serious effects on the environment. Traffic noise, construction sites, airports and other factories produce extremely high amounts of noise. Noise-related issues include diseases linked to stress, high blood pressure, speech obscuration, hearing loss, insomnia and decreased productivity. People will be able to live healthier lives if the environment is preserved and kept clean.

Big data and real-time processing challenges are not addressed by prior research on air pollution. The amount of data increases as the globe develops, making management and preparation challenging. In order to boost processing efficiency while improving accuracy and handling huge data, Apache Spark is implemented in this work. Big data workloads are processed with the open-source distributed processing system Apache Spark. For quick analytical queries against any quantity of data, it uses in-memory caching and efficient query execution. Time series forecasting using Apache Spark enables the monitoring, tracking, and prediction of the behaviour of pollution data over a range of years.

The primary goal of the research is to determine the many forms of pollution and their effects on the actual land where people live. In this project, two distinct types of pollution are analysed, and it projects how they would affect society if no action is made to thwart them from ensuing in the future. Every member of society must be aware of the environment's inevitable pollution as a result of modern technology and take action to restore the earth to its natural state of green. According to studies using sophisticated modelling, it would take between 3 and 5 million years without human intervention to restore a world free of pollution, and if this trend continued, it would take 4 billion years for the pollution to be completely eradicated.

CHAPTER 2

LITERATURE SURVEY

2.1 Air Quality Quantification in Taiwan Using Machine Learning Techniques in Apache Spark Platform

Kandath, Sreenand

The air quality is really poor, and the oxygen is sold in bottles. The machine learning algorithm is effectively implemented to accurately calculate the Air Quality Index (AQI). Apache Spark is used to perform the function on the historical air quality data of the island of Taiwan. The boosted tree was found to be the best model for their research

2.2 Analysis and Prediction of Air Pollutant Using Machine Learning

Chalumuru Suresh, B. V. Kiranmayee, Balannolla Sneha

Air pollution has risen drastically, and it affects health issues in the worst manner. People all around the world suffer from diseases related to the lungs, respiration, cardiovascular illness, etc. The estimation of pollutant concentrations in the air we inhale will help to aid the government and give awareness to different age groups of people. Their research aim was to find the model that predicts the exact concentration in the atmosphere.

2.3 Time series analysis and forecasting of air quality index

M. C. Lineesh

The most essential element for sustaining human existence is thought to be air. It is important to remember that even a slight change in the air's chemical makeup will have a big influence on the ecological balance of our planet. Their work contained PM10 level analysis in the Kozhikode district using a Self-Exciting Threshold Autoregressive (SETAR) model during COVID-19.

2.4 Analysis of Air Pollution in Three Cities of Kerala by Using Air Quality Index

S.N Jyothi, Kishan Kartha, Divesh, Adarsh Mohan, Jithin Pai U, Geena Prasad

Air pollution has a direct effect on both the globe and the planet. Here, air pollution is measured using the parameter Air Quality Index (AQI). This research paper collected data from six major locations across Kerala and compared them. Here the AQI is calculated and plotted against the month for individual locations.

2.5 Compositional time series analysis for Air Pollution Index data

Nasr Ahmed AL-Dhurafi, Nurulkamal Masseran , Zamira Hasanah Zamzuri

This research focused on the Air Pollution Index (API) and its five major contributing pollutants. Here, they proved that API can be shown as compositional data. The data collected by API is demonstrated in Klang, Malaysia, for the period of January 2005 to December 2014. A best-fitted model is a VAR model with zero trends, which is used for forecasting the upcoming 12 months.

2.6 Predictive mapping of urban air pollution using Apache Spark on a Hadoop cluster

Marjan Asgari, Mahdi Farnaghi, Zeinab Ghaemi

Air pollution is a serious environmental problem in densely populated areas. The maps of air pollution over the next 24 hours in this article were predicted using distributed processing methods. Tehran, the capital of Iran, is chosen as the site for the research. In this case, on Apache Spark, the Hadoop cluster serves as the framework. The Inverse Distance Weighting (IDW) approach is employed to distribute the anticipated results over the whole city.

2.7 A Scalable and Reliable Model for Real-time Air Quality Prediction

Liying Li, Zhi Li, Lara Reichmann, Diane Woodbridge

Californian datasets on air quality were gathered for this study. Machine learning models are applied to a pipeline that stores, processes, and makes predictions. Using the Spark machine learning and Apache Spark SQL libraries, the pipeline was created. By taking into account a 10-year dataset, a prediction model was created using logistic regression and random forest classification.

2.8 Analysis of Noise Pollution during Dussehra Festival in Bhubaneswar Smart City in India: A Study Using Machine Intelligence Models

Sourav Kumar Bhoi, Chittaranjan Mallick, Chitta Ranjan Mohanty, and Ranjan Soumya Nayak

Noise pollution has increased as a result of increased urbanization. Data from the State Pollution Control Board in Odisha, India, for the years 2015 to 2020 was used to estimate the noise level in Bhubaneswar city during Dussehra in the year 2020. An analytics tool called Orange 3.26 has been used for a variety of supervised models. It is shown that DT and RF have the highest prediction accuracy, with a classification accuracy of 0.925.

2.9 A general procedure to generate models for urban environmental-noise pollution using feature selection and machine learning methods

Antonio J.Torija, Diego P. Ruiz

Due to the numerous relationships between the variables, predicting environmental noise is a challenging, non-linear task. Here, the issue is resolved by using a system based on feature-selection approaches and regression machine-learning techniques. Both feature selection and regression are approached separately using three different methods. Among all the models, LAeq prediction models are the most complex and time-consuming.

2.10 Traffic noise prediction model of an Indian road: an increased scenario of vehicles and honking

Chaitanya Thakre, Vijaya Laxmi, Ritesh Vijay, Deepak J. Killedar, Rakesh Kumar

As compared to other pollutants, noise has significant negative consequences on health. This analysis used information from Nagpur's secondary roads during the years 2012 and 2019. Multiple regression was used to construct a noise prediction model, and it was discovered that there is an increase in sound pressure due to traffic and honking

2.11 Noise Pollution and Urban Planning

Juan Miguel Barrión Morillas, Guillermo Rey Gozalo, David Montes González, Pedro Atanasio Moraga, Rosendo Vílcez-Gómez

The relationship between noise levels and urban planning and morphology has been explored in several studies in the recent past. There have been studies of the city's functional aspects, green spaces, parking areas, lane numbers, traffic lights, as well as the surrounding urban environment. Traditional places might not be able to make all the required modifications in order to meet acceptable noise pollution standards. Those responsible for redesigning urban designs can take into account noise-determining elements that may result in more successful urban areas.

2.12 Noise prediction of chemical industry park based on multi-station Prophet and multivariate LSTM fitting model

Qingtian Zeng, Yu Liang, Geng Chen, Hua Duan & Chunguo Li

This research uses noise data that was collected in Chemical Park. The integration of data from the wind speed and vehicle flow is done using a proposed LSTM model. After several comparisons, the multi-PL model prediction approach is found to be more accurate and stable.

CHAPTER 3

PROPOSED SOLUTION

3.1 METHODOLOGY:

3.1.1 TRAINING FLOW:

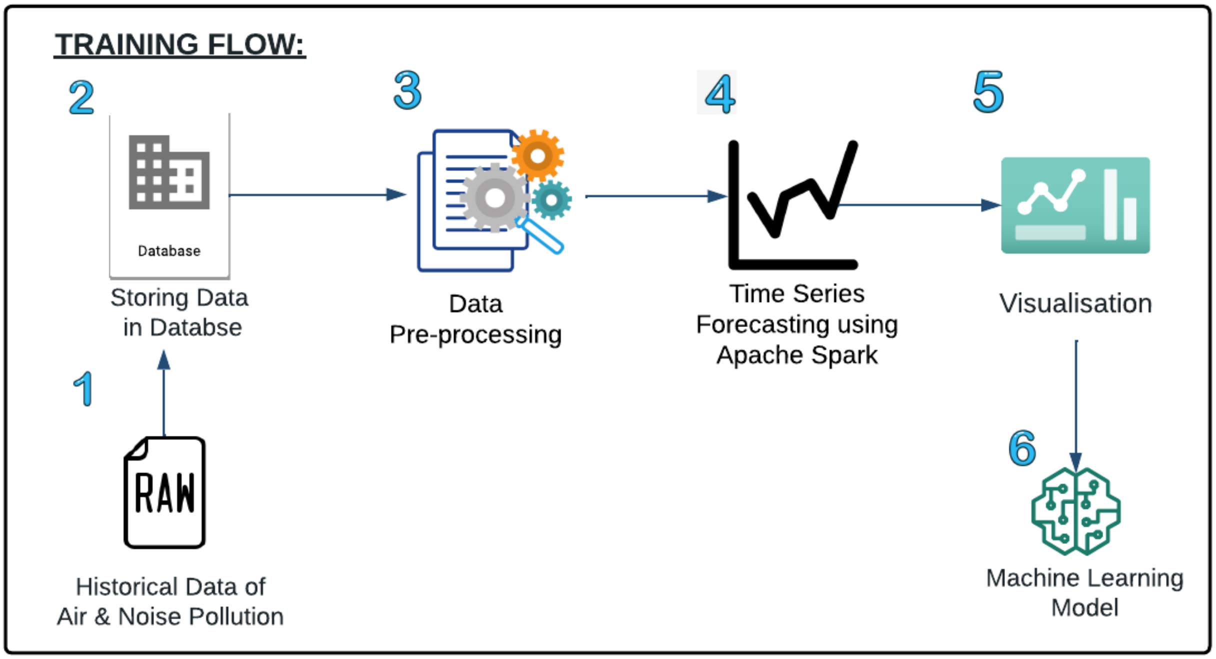


Figure 3.1 Diagrammatic Representation of Training Flow and Prediction Flow

The process's first step is the collection of data. The data is collected from the World Health Organization (WHO). Two different types of datasets, that is air pollution and noise pollution data are collected. The dataset comprises features like the pollution rate, the geographical location, and other features which affects the pollution rate. This raw data contains a lot of anomalies, hence the pre-processing steps like handling missing values, handling outliers, encoding the string values, and feature scaling are explicitly performed. Next, the cleaned data is directly equipped into the PySpark Machine Learning model and the Time Series Forecasting is performed. The visualisation is made with Google Studios. The estimation and the trend are viewed for each location for the upcoming years in accordance with their pollutant rate. A separate model for both air pollution and noise pollution is created. These models are stored so that they can be used when we want to perform time series forecasting in spark for air pollution and noise pollution.

3.1.2 PREDICTION FLOW:

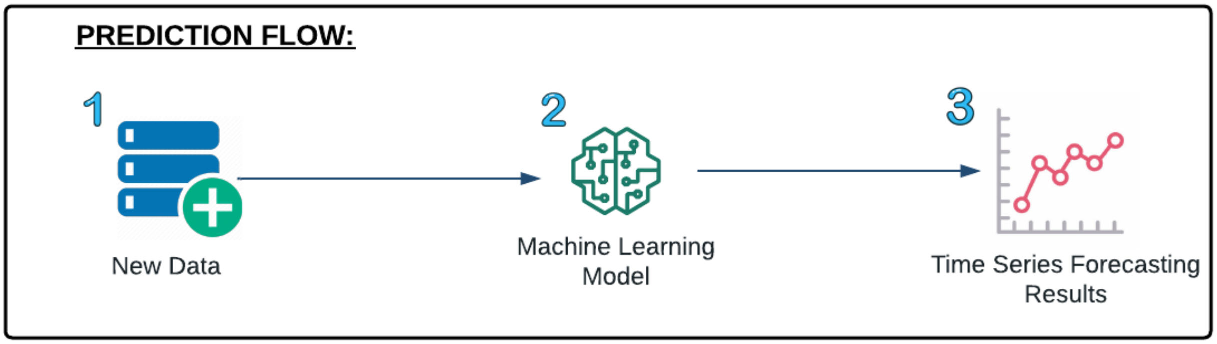


Figure 3.2 Diagrammatic Representation of Prediction Flow

Now, when the new is collected the data is provided to the ML model that is created with PySpark for the time series forecasting. The preprocessing steps, training, and prediction are made in this way.

3.2 TECH STACK

APACHE SPARK (PySpark):

Apache Spark is an open-source processing system for big data workloads. Apache Spark is a multi-language engine for executing machine learning on a user-friendly type either by single-node machines or clusters. Its execution is fast. It supports development in Java, Scala, Python and R. We have used this Apache Spark with python, it is also called PySpark. It also has fault tolerance capacity.

STREAMLIT:

Streamlit is an open-source application for easily building web applications. For using this there is no prior knowledge required in the front-end website development. The charts can also be displayed easily.

GOOGLE DATA STUDIO:

Making data in an understandable manner regardless of people’s knowledge is an art. Google data studio is being used to visualise all the result content. This provides user interactive visualisation.

JUPYTER LAB:

JupyterLab is a web-based interactive computing platform. This makes users work easily on machine-learning type of tasks. A user-friendly environment is provided in this JupyterLab. We implemented all the PySpark code using this.

MACHINE LEARNING LIBRARIES:

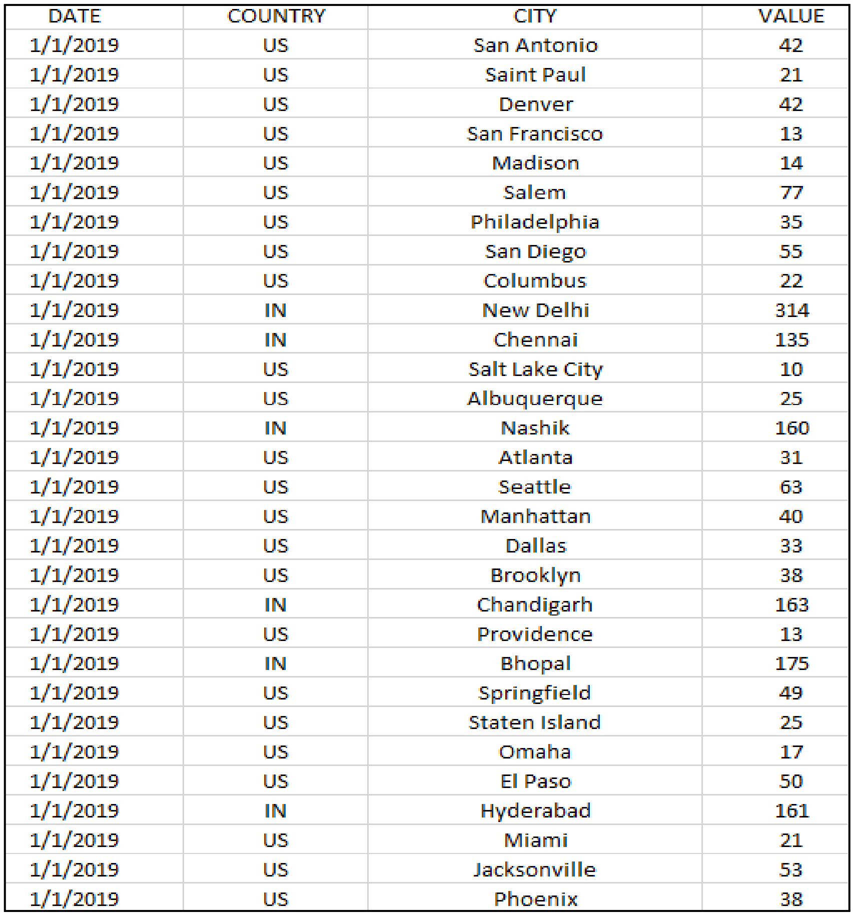
PANDAS:

Pandas is a python library used in data science for performing tasks related to machine learning. It has the feature to work with any type of data.

NUMPY:

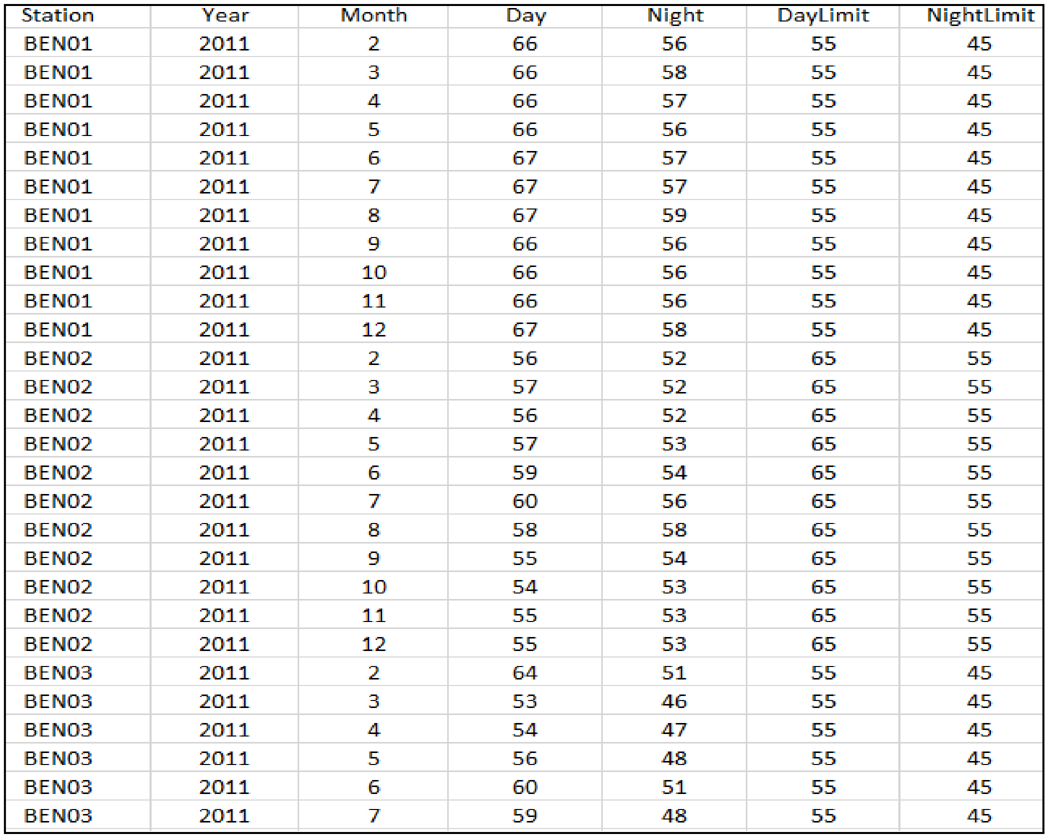
NumPy is an open source for majorly working with numeric data. NumPy is used for working with multi-dimensional arrays. It is faster than traditional Python.

3.3 DATASET:



DATE	COUNTRY	CITY	VALUE
1/1/2019	US	San Antonio	42
1/1/2019	US	Saint Paul	21
1/1/2019	US	Denver	42
1/1/2019	US	San Francisco	13
1/1/2019	US	Madison	14
1/1/2019	US	Salem	77
1/1/2019	US	Philadelphia	35
1/1/2019	US	San Diego	55
1/1/2019	US	Columbus	22
1/1/2019	IN	New Delhi	314
1/1/2019	IN	Chennai	135
1/1/2019	US	Salt Lake City	10
1/1/2019	US	Albuquerque	25
1/1/2019	IN	Nashik	160
1/1/2019	US	Atlanta	31
1/1/2019	US	Seattle	63
1/1/2019	US	Manhattan	40
1/1/2019	US	Dallas	33
1/1/2019	US	Brooklyn	38
1/1/2019	IN	Chandigarh	163
1/1/2019	US	Providence	13
1/1/2019	IN	Bhopal	175
1/1/2019	US	Springfield	49
1/1/2019	US	Staten Island	25
1/1/2019	US	Omaha	17
1/1/2019	US	El Paso	50
1/1/2019	IN	Hyderabad	161
1/1/2019	US	Miami	21
1/1/2019	US	Jacksonville	53
1/1/2019	US	Phoenix	38

Figure 3.3 Dataset Utilised for Air Quality Index Estimation



Station	Year	Month	Day	Night	DayLimit	NightLimit
BEN01	2011	2	66	56	55	45
BEN01	2011	3	66	58	55	45
BEN01	2011	4	66	57	55	45
BEN01	2011	5	66	56	55	45
BEN01	2011	6	67	57	55	45
BEN01	2011	7	67	57	55	45
BEN01	2011	8	67	59	55	45
BEN01	2011	9	66	56	55	45
BEN01	2011	10	66	56	55	45
BEN01	2011	11	66	56	55	45
BEN01	2011	12	67	58	55	45
BEN02	2011	2	56	52	65	55
BEN02	2011	3	57	52	65	55
BEN02	2011	4	56	52	65	55
BEN02	2011	5	57	53	65	55
BEN02	2011	6	59	54	65	55
BEN02	2011	7	60	56	65	55
BEN02	2011	8	58	58	65	55
BEN02	2011	9	55	54	65	55
BEN02	2011	10	54	53	65	55
BEN02	2011	11	55	53	65	55
BEN02	2011	12	55	53	65	55
BEN03	2011	2	64	51	55	45
BEN03	2011	3	53	46	55	45
BEN03	2011	4	54	47	55	45
BEN03	2011	5	56	48	55	45
BEN03	2011	6	60	51	55	45
BEN03	2011	7	59	48	55	45

Figure 3.4 Dataset Utilised for Noise Pollution Rate Estimation

The datasets are obtained from the World Health Organization (WHO). Figure 3.1 shows the data on the air quality index in the US and India from 2019 to 2021. Figure 3.2 shows the data on noise pollution values at several stations in India from 2011 to 2018.

3.4 DEVELOPMENT ENVIRONMENT

3.4.1 Hardware Requirements

- Processor: intel i7 processor
- Motherboard: intel chipset motherboard
- Ram: 8GB or more
- Cache: 512kb
- Hard disk: 16GB hard disk recommended
- Speed: 2.7ghz and more

3.4.2 Software Requirements

S.No	Software	Version	URL
1	Operating system	Ubuntu 22.04.1	https://ubuntu.com/download/desktop
2	Microsoft Excel	Microsoft 365	https://www.microsoft.com/en-in/microsoft-365/excel
3	Python	3.9.10	https://www.python.org/
4	VSCode	1.73	https://code.visualstudio.com/download
5	PySpark	3.3.1	https://spark.apache.org/
6	Pandas	1.5.2	https://pandas.pydata.org/
7	NumPy	1.23.5	https://pypi.org/project/numpy/
8	Jupyter Lab	3.4.0	https://jupyter.org/
9	Google Data Studio	Oct 7, 2021	https://datastudio.google.com/u/0/
11	Sklearn	1.1.3	https://pypi.org/project/scikit-learn/

CHAPTER 4

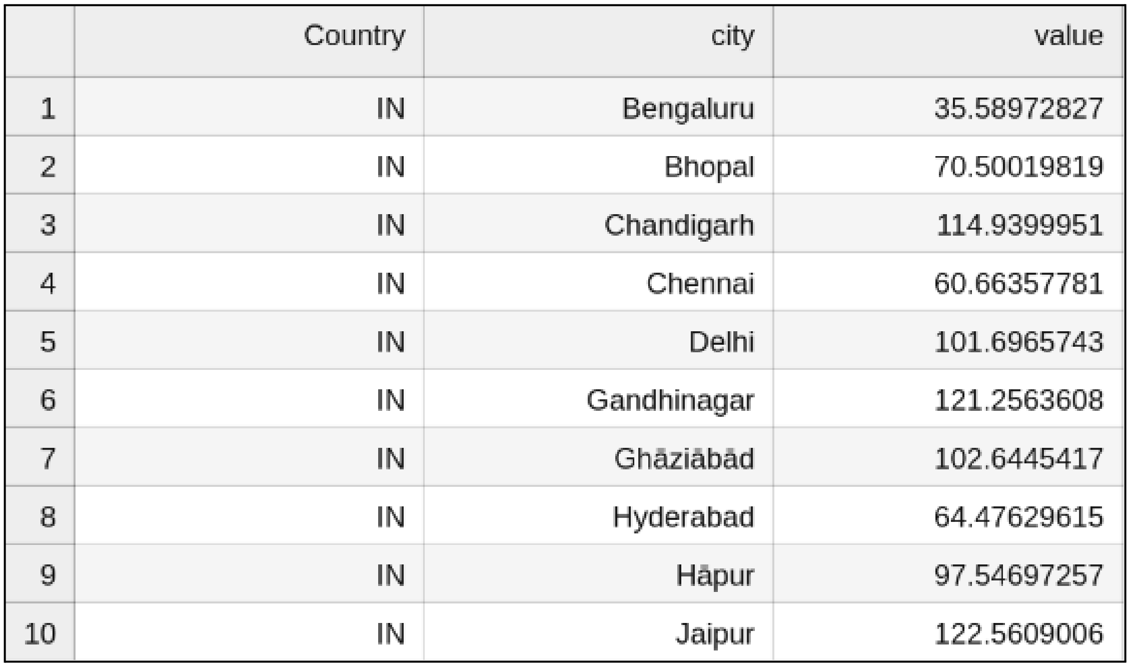
IMPLEMENTATION USING PySpark

4.1 IMPLEMENTATION OF AIR POLLUTION

4.1.1 Air Pollution PySparkCode

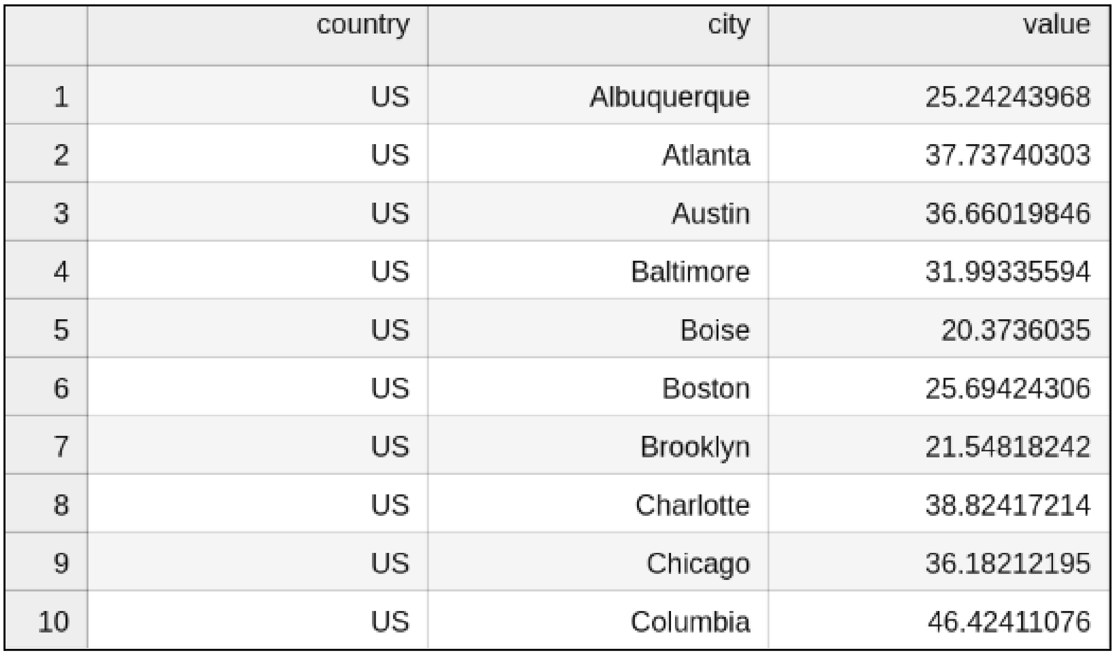
```
from pyspark.sql.functions import pandas_udf, PandasUDFType
from pyspark.sql.types import
StructType,StructField,StringType,LongType,DoubleType,FloatType

import statsmodels.tsa.api as sm
import numpy as np
import pandas as pd

from pyspark.sql import SparkSession

spark = SparkSession.builder.appName('TSF-pollution').getOrCreate()


data = spark.read.format('csv').options(header='true',
inferSchema='true').load('dataset/air_quality_index.csv').select('DATE',
'COUNTRY','CITY','VALUE')

schema = StructType([StructField('COUNTRY', StringType(), True),
                     StructField('CITY', StringType(), True),
                     StructField('VALUE', DoubleType(), True)])

selected_com = data.groupBy(['COUNTRY','CITY']).count().filter("count >
104").select("COUNTRY","CITY")
data_selected_store_departments = data.join(selected_com,
['COUNTRY','CITY'],'inner')

# df.set_index('DATE',inplace = True)

@pandas_udf(schema, PandasUDFType.GROUPED_MAP)


def holt_winters_time_series_udf(data):

    data.set_index('DATE',inplace = True)
    time_series_data = data['VALUE']

    ##the model
    model_monthly =
sm.ExponentialSmoothing(np.asarray(time_series_data),trend='add').fit()

    ##forecast values
    forecast_values = pd.Series(model_monthly.forecast(1),name =
'fitted_values')

    return pd.DataFrame({'COUNTRY': [str(data.COUNTRY.iloc[0])],'CITY':
[str(data.CITY.iloc[1])],'VALUE': [forecast_values[0]]})

forecasted_spark_df =
data.groupby(['COUNTRY','CITY']).apply(holt_winters_time_series_udf)

forecasted_spark_df.show(10)

print((forecasted_spark_df.count(), len(forecasted_spark_df.columns)))
```

4.1.2 Output of Air pollution Estimation



	Country	city	value
1	IN	Bengaluru	35.58972827
2	IN	Bhopal	70.50019819
3	IN	Chandigarh	114.9399951
4	IN	Chennai	60.66357781
5	IN	Delhi	101.6965743
6	IN	Gandhinagar	121.2563608
7	IN	Ghāziābād	102.6445417
8	IN	Hyderabad	64.47629615
9	IN	Hāpur	97.54697257
10	IN	Jaipur	122.5609006

Figure 4.1 Output image shows the air pollution in India



	country	city	value
1	US	Albuquerque	25.24243968
2	US	Atlanta	37.73740303
3	US	Austin	36.66019846
4	US	Baltimore	31.99335594
5	US	Boise	20.3736035
6	US	Boston	25.69424306
7	US	Brooklyn	21.54818242
8	US	Charlotte	38.82417214
9	US	Chicago	36.18212195
10	US	Columbia	46.42411076

Figure 4.2 Output image shows the air pollution in the United States

4.2 IMPLEMENTATION OF NOISE POLLUTION

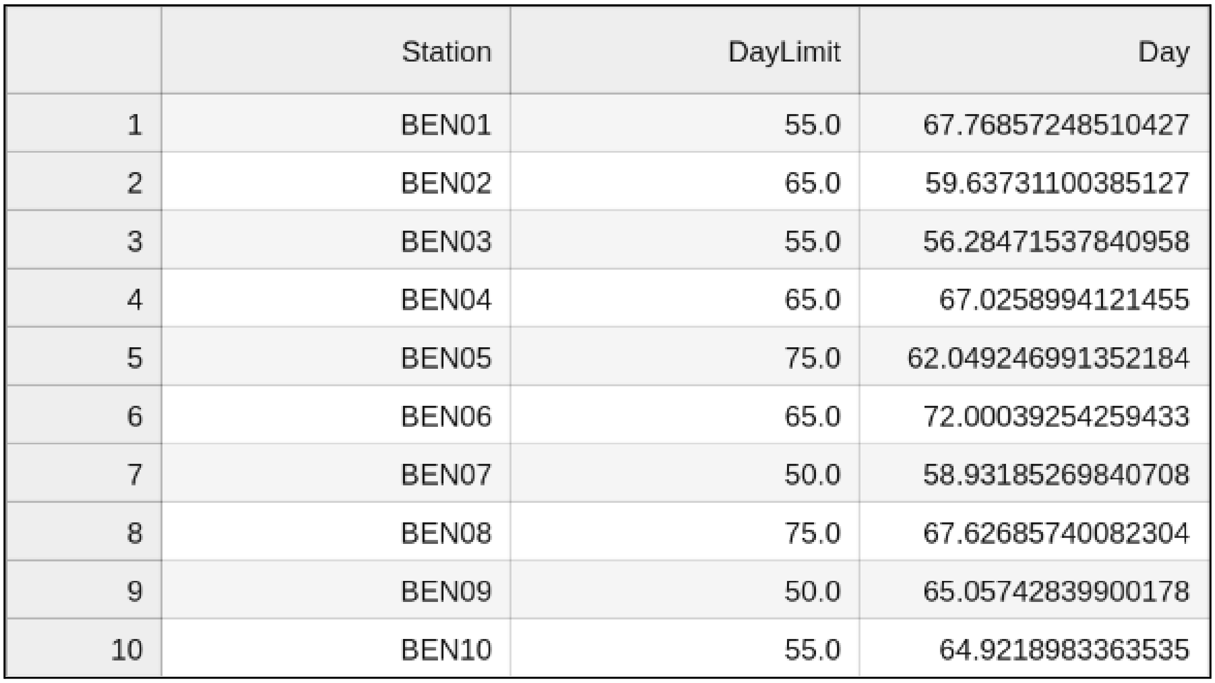
4.2.1 Noise Pollution Variations At Daytime

```
from pyspark.sql.functions import pandas_udf, PandasUDFType
from pyspark.sql.types import StructType, StructField, StringType, LongType,
DoubleType, FloatType

import statsmodels.tsa.api as sm
import numpy as np
import pandas as pd

from pyspark.sql import SparkSession

import warnings
warnings.filterwarnings('ignore')

spark = SparkSession.builder.appName('TSF-noise-pollution').getOrCreate()

# df =
pd.read_csv("/Users/sanjju/projects/datasets/noise-pollution/station_month.c
sv")
```

```
data = spark.read.format('csv').options(header='true',
inferSchema='true').load('dataset/station_month.csv').select('Year',
'Station', 'DayLimit', 'Day')

schema = StructType([StructField('Station', StringType(), True),
                    StructField('DayLimit', DoubleType(), True),
                    StructField('Day', DoubleType(), True)])

selected_com = data.groupBy(['Station',
'DayLimit']).count().filter("DayLimit > 0").select("Station","DayLimit")
data_selected_store_departments = data.join(selected_com, ['Station',
'DayLimit'], 'inner')

# selected_com = data['Station']
# data_selected_store_departments = data['Station']

# df.set_index('Year',inplace = True)

@pandas_udf(schema, PandasUDFType.GROUPED_MAP)

def holt_winters_time_series_udf(data):

    data.set_index('Year',inplace = True)
    time_series_data = data['Day']

    ##the model
    model_monthly =
sm.ExponentialSmoothing(np.asarray(time_series_data),trend='add').fit()

    ##forecast values
    forecast_values = pd.Series(model_monthly.forecast(1),name =
'fitted_values')

    return pd.DataFrame({'Station': [str(data.Station.iloc[0])], 'DayLimit':
[int(data.DayLimit.iloc[1])], 'Day': [forecast_values[0]]})

forecasted_spark_df = data_selected_store_departments.groupby(['Station',
'DayLimit']).apply(holt_winters_time_series_udf)


## to see the forecasted results
forecasted_spark_df.show(10)

print((forecasted_spark_df.count(), len(forecasted_spark_df.columns)))
```

```
#
forecasted_spark_df.write.csv('dataset/forecasted-noise-pollution-day.csv')
```

4.2.2 Output of Noise Variations at Day time:



	Station	DayLimit	Day
1	BEN01	55.0	67.76857248510427
2	BEN02	65.0	59.63731100385127
3	BEN03	55.0	56.28471537840958
4	BEN04	65.0	67.0258994121455
5	BEN05	75.0	62.049246991352184
6	BEN06	65.0	72.00039254259433
7	BEN07	50.0	58.93185269840708
8	BEN08	75.0	67.62685740082304
9	BEN09	50.0	65.05742839900178
10	BEN10	55.0	64.9218983363535

Figure 4.3 Output image shows the Noise pollution at various stations during Day time

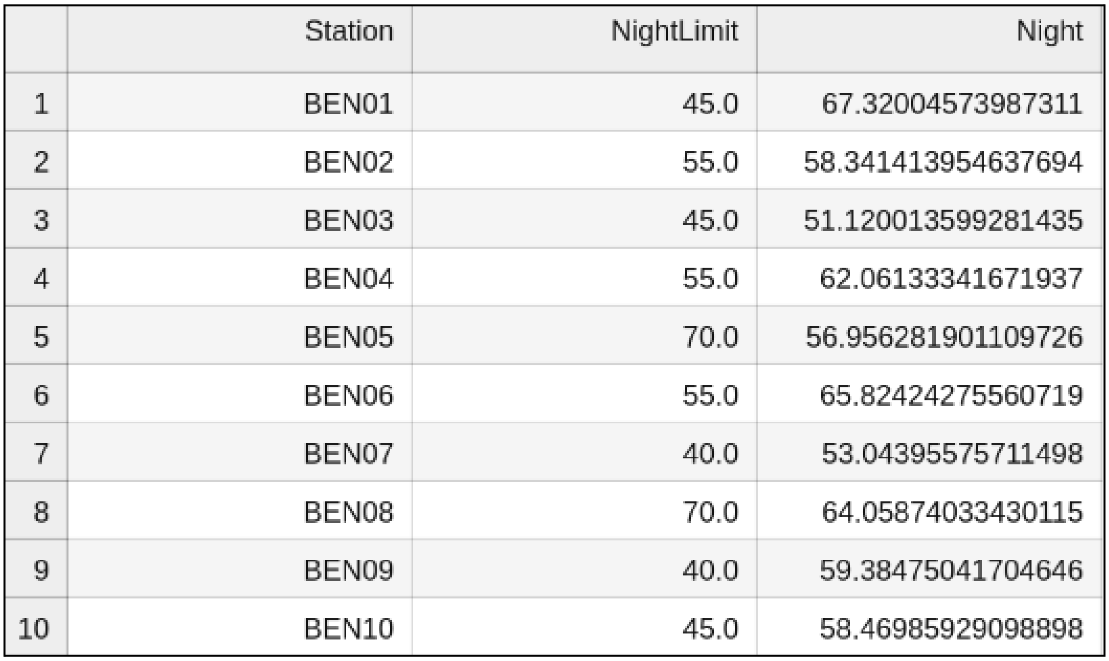
4.2.3 Noise Pollution Variations At Night time

```
from pyspark.sql.functions import pandas_udf, PandasUDFType
from pyspark.sql.types import
StructType,StructField,StringType,LongType,DoubleType,FloatType

import statsmodels.tsa.api as sm
import numpy as np
import pandas as pd

from pyspark.sql import SparkSession
```

```
import warnings
warnings.filterwarnings('ignore')

spark = SparkSession.builder.appName('TSF-noise-pollution').getOrCreate()


data = spark.read.format('csv').options(header='true',
inferSchema='true').load('dataset/station_month.csv').select('Year',
'Station', 'NightLimit', 'Night')

schema = StructType([StructField('Station', StringType(), True),
                     StructField('NightLimit', DoubleType(), True),
                     StructField('Night', DoubleType(), True)])

selected_com =
data.groupBy(['Station','NightLimit']).count().filter("NightLimit >
0").select("Station","NightLimit")
data_selected_store_departments = data.join(selected_com,
['Station','NightLimit'],'inner')


@pandas_udf(schema, PandasUDFType.GROUPED_MAP)

def holt_winters_time_series_udf(data):

    data.set_index('Year',inplace = True)
    time_series_data = data['Night']

    ##the model
    model_monthly =
sm.ExponentialSmoothing(np.asarray(time_series_data),trend='add').fit()

    ##forecast values
    forecast_values = pd.Series(model_monthly.forecast(1),name =
'fitted_values')

    return pd.DataFrame({'Station': [str(data.Station.iloc[0])],
'NightLimit': [int(data.NightLimit.iloc[1])], 'Night':
[forecast_values[0]]})

forecasted_spark_df = data_selected_store_departments.groupby(['Station',
'NightLimit']).apply(holt_winters_time_series_udf)


## to see the forecasted results
```

```
forecasted_spark_df.show(10)

print((forecasted_spark_df.count(), len(forecasted_spark_df.columns)))

#
forecasted_spark_df.write.csv('dataset/forecasted-noise-pollution-night.csv'
)
```

4.2.4 Output of Noise Variations at Night time:



	Station	NightLimit	Night
1	BEN01	45.0	67.32004573987311
2	BEN02	55.0	58.341413954637694
3	BEN03	45.0	51.120013599281435
4	BEN04	55.0	62.06133341671937
5	BEN05	70.0	56.956281901109726
6	BEN06	55.0	65.82424275560719
7	BEN07	40.0	53.04395575711498
8	BEN08	70.0	64.05874033430115
9	BEN09	40.0	59.38475041704646
10	BEN10	45.0	58.46985929098898

Figure 4.4 Output image shows the Noise pollution at various stations during Night time

CHAPTER 5

RESULT AND ANALYSIS

5.1 RESULT AND ANALYSIS OF AIR POLLUTION:

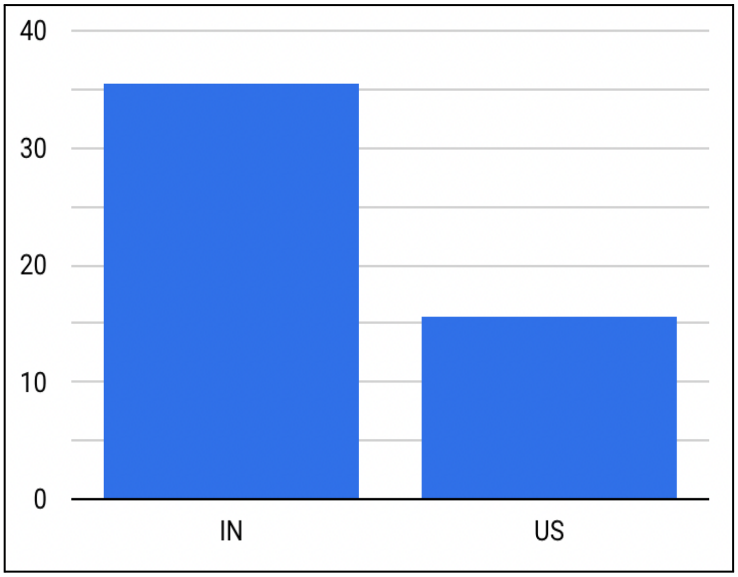


Figure 5.1 AQI (Air Quality Index) Comparison for India and United States

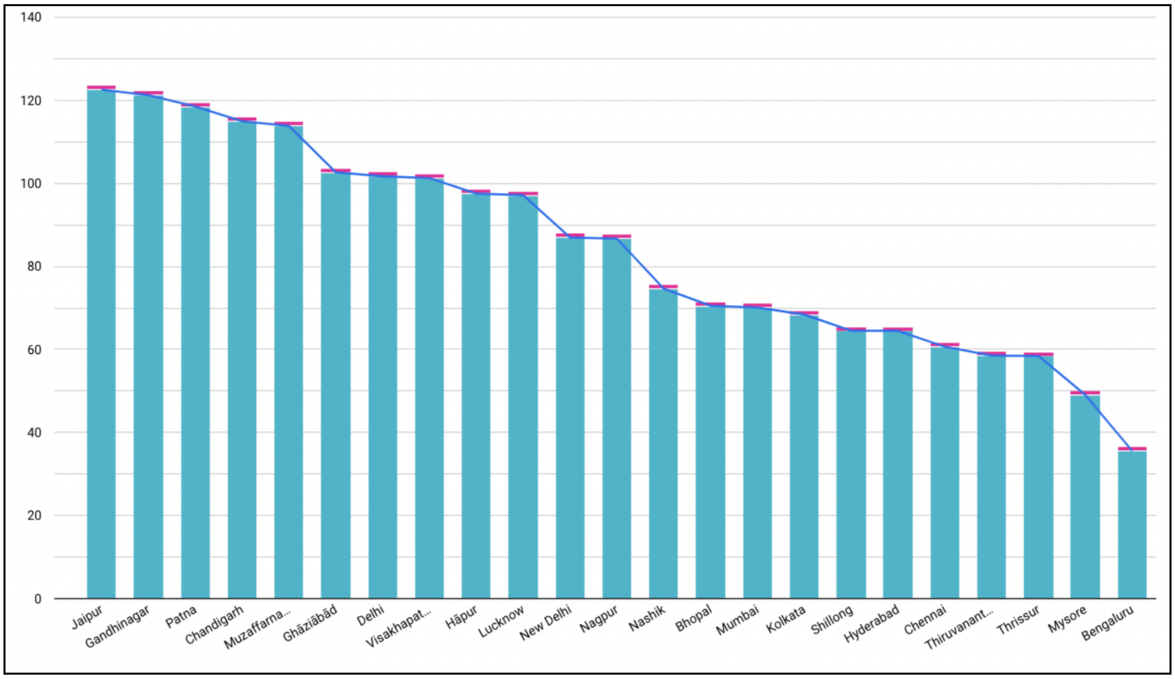


Figure 5.2 Indian Air Pollution Time Series for December 2022
X-Axis: Indian Cities
Y-Axis: Air Quality Index(AQI) for December 2022

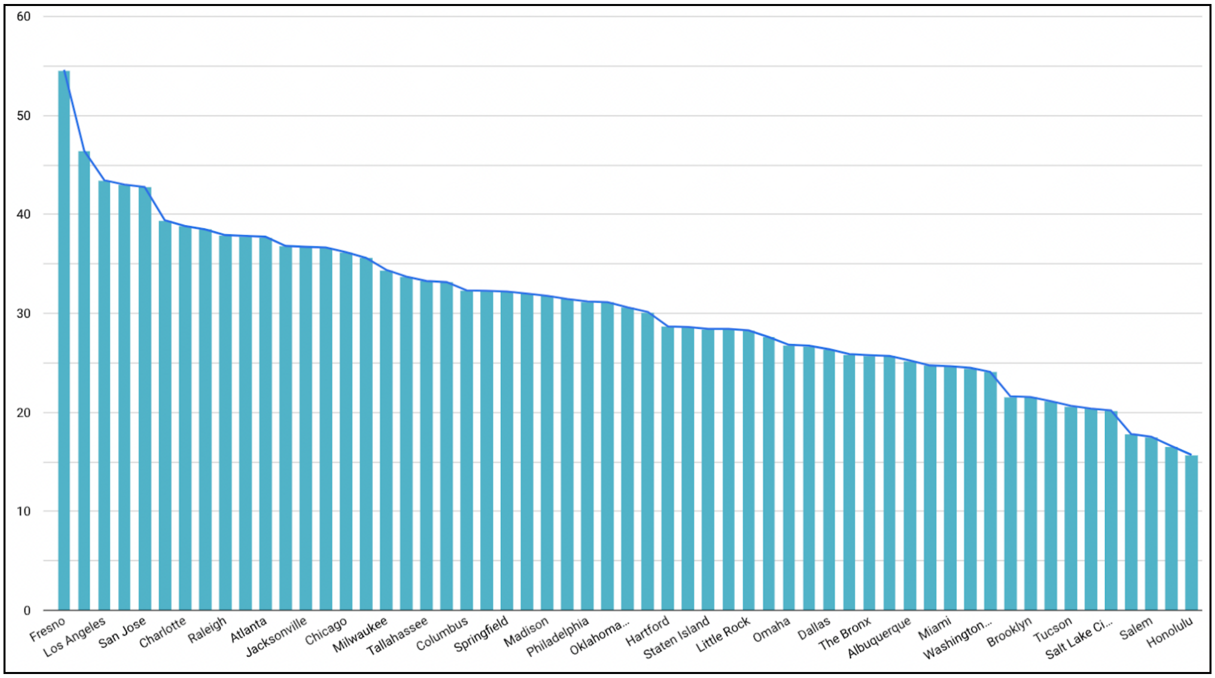


Figure 5.3 United States's Air Pollution Index of Time Series for December 2022
X-Axis: US Cities
Y-Axis: Air Quality Index(AQI) for December 2022

5.2 RESULT AND ANALYSIS OF NOISE POLLUTION:

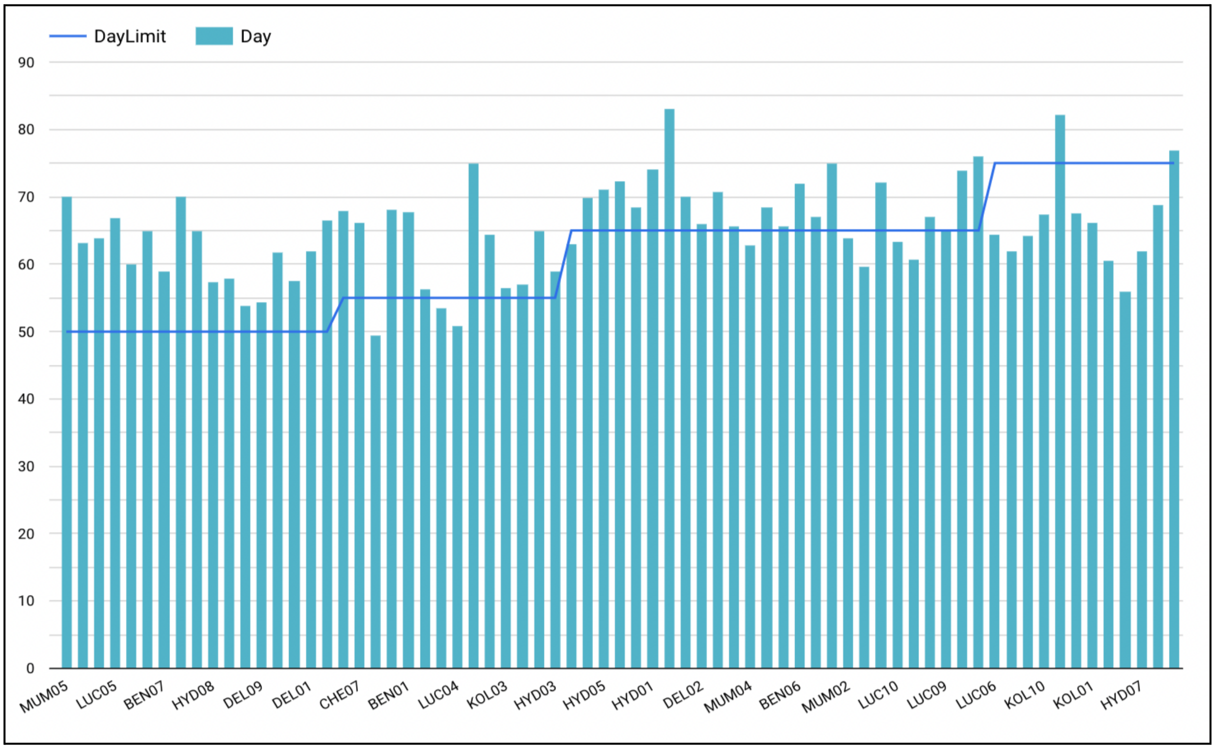


Figure 5.4 Time Series Forecasting for Noise Pollution during the day of December 2022
X-Axis: Indian City Station Code
Y-Axis: Noise Pollution rate during daytime

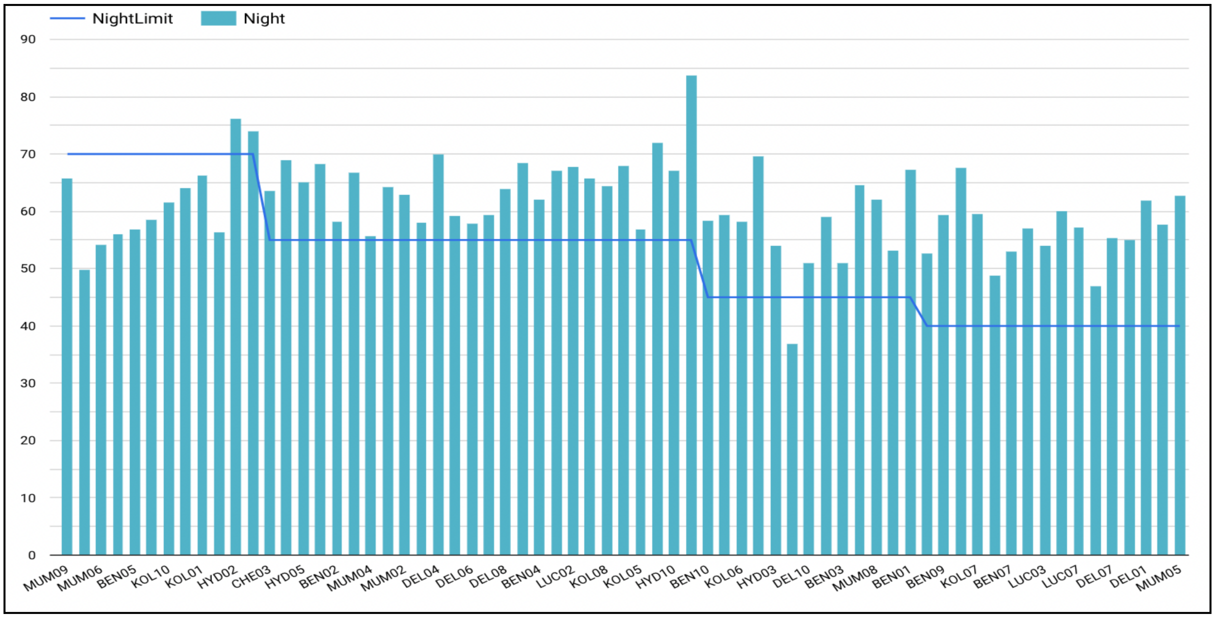


Figure 5.5 Time Series Forecasting for noise pollution during the night of December 2022
X-Axis: Indian City Station Code
Y-Axis: Noise Pollution rate during night-time

CHAPTER 6

CONCLUSION

CONCLUSION:

Air pollution and noise pollution constitute major pollutants that are detrimental to human health and the planet as a whole. Here time series is employed with Apache Spark for its quick processing time and dynamic nature. The everyday levels of noise and air pollution in India and the US during December 2022 are depicted using visualization techniques. The ratio between the US and Indian state air quality indexes has been found to be 7:3.

The outcomes can be used in a variety of fields to preserve a sustainable environment. Future improvements will recommend preventative measures that should be taken in a certain region along with pollutant levels.

REFERENCES:

[1] Kandath, Sreenand (2019) Air Quality Quantification in Taiwan Using Machine Learning Techniques in Apache Spark Platform. Masters thesis, Dublin, National College of Ireland.

[2] Chalumuru Suresh, B. V. Kiranmayee, Balannolla Sneha. (2022). Analysis and Prediction of Air Pollutant Using Machine Learning. In: Reddy, A.B., Kiranmayee, B., Mukkamala, R.R., Srujan Raju, K. (eds) Proceedings of Second International Conference on Advances in Computer Engineering and Communication Systems. Algorithms for Intelligent Systems. Springer, Singapore.

[3] M. C. Lineesh. (2021). Time series analysis and forecasting of air quality index. INTERNATIONAL CONFERENCE ON COMPUTATIONAL SCIENCES-MODELLING, COMPUTING AND SOFT COMPUTING (CSMCS 2020).

[4] S.N Jyothi, Kishan Kartha, Divesh, Adarsh Mohan, Jithin Pai U, Geena Prasad. (2019). Analysis of Air Pollution in Three Cities of Kerala by Using Air Quality Index. Journal of Physics: Conference Series, 1362(1), 012110.

[5] Nasr Ahmed AL-Dhurafi, Nurulkamal Masseran, Zamira Hasanah Zamzuri. (2018). Compositional time series analysis for Air Pollution Index data. Stochastic Environmental Research and Risk Assessment.

[6] Marjan Asgari, Mahdi Farnaghi, Zeinab Ghaemi. (2017). Predictive mapping of urban air pollution using Apache Spark on a Hadoop cluster.

[7] Liying Li, Zhi Li, Lara Reichmann, Diane Woodbridge. (2019). A Scalable and Reliable Model for Real-time Air Quality Prediction. IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI).

[8] Sourav Kumar Bhoi, Chittaranjan Mallick, Chitta Ranjan Mohanty, and Ranjan Soumya Nayak. (2022). Analysis of Noise Pollution during Dussehra Festival in Bhubaneswar Smart City in India: A Study Using Machine Intelligence Models.

[9] Antonio J.Torija, Diego P. Ruiz. (2015). A general procedure to generate models for urban environmental-noise pollution using feature selection and machine learning methods. Science of The Total Environment, 505, 680–693.

[10] Chaitanya Thakre, Vijaya Laxmi, Ritesh Vijay, Deepak J. Killedar, Rakesh Kumar. (2020). Traffic noise prediction model of an Indian road: an increased scenario of vehicles and honking. Environmental Science and Pollution Research.

[11] Morillas, J. M. B., Gozalo, G. R., González, D. M., Moraga, P. A., & Vílchez-Gómez, R. (2018). Noise Pollution and Urban Planning. Current Pollution Reports, 4(3), 208–219.

[12] Qingtian Zeng, Yu Liang, Geng Chen, Hua Duan, Chunguo Li. Noise prediction of chemical industry park based on multi-station Prophet and multivariate LSTM fitting model. *EURASIP J. Adv. Signal Process.* 2021, 106 (2021).