ANALYSIS OF WATER RESOURCES AND WATER MANAGEMENT FOR CHENNAI CITY

CSE3020 – DATA VISUALISATION

PROJECT BASED COMPONENT REPORT

By

TANYA VATS **-18BCB0031**

NITIN PRAMOD RANJAN**-18BCE0272**

SIDDHARTH GARG**- 18BCE0525**

AMAN KUMAR**- 18BCE0659**

MANAV JAIN**-18BCE0695**

Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

June, 2021

DECLARATION

I hereby declare that the report entitled "Analysis of water resources and water management for Chennai City" submitted by us, for the CSE3020 DATA VISUALISATION (EPJ) to VIT is a record of bonafide work carried out by me under the supervision of Dr.S.VENGADESWARAN .

I further declare that the work reported in this report has not been submitted and will not be submitted, either in part or in full, for any other courses in this institute or any other institute or university.

Place: Vellore

Date: 5-06-2021

Signature of the Candidate

TANYA VATS

NITIN PRAMOD RANJAN

SIDDHARTH GARG

AMAN KUMAR

MANAV JAIN

CONTENTS

ABSTRACT

Technological advances are driving exponential growth in data, improving the efficiency of many sectors and disrupting others. Extracting meaningful information quickly from the modern-day deluge of data is now a requirement for being successful. However, generating decision-ready data or predicting future outcomes from environmental or operational data is challenging and requires specialized skills. Research teams can tap into big data insights through the inclusion of data scientists — individuals with interdisciplinary backgrounds that include data analysis, statistics, data visualization, computer science, and mathematics.

Data visualization is a valuable tool for making high-volume, complex data accessible to stakeholders, policy makers, and managers to facilitate data-driven conversations about environmental issues.

Through this project, we aim to analyse the historical data about water resources available for Chennai city. With the help of visualizations and statistical modelling techniques we will be able to predict whether the currently available water resources will be able to meet the water requirements of the residents of Chennai and surrounding areas.

We used Exploratory data analysis and ARIMA model to do the Data Visualization, prediction and dashboard implementation with the help of Tableau and Google colab.

INTRODUCTION

Chennai, formerly known as Madras, is the capital of Tamil Nadu, a state the eastern coast of India. Located on the Coromandel Coast off the Bay of Bengal, the city of Chennai is one of the largest cultural, economic and educational centres of south India. It is also one among the 100 cities to be developed as a smart city under the Smart Cities Mission launched by PM Narendra Modi in 2015. However, in spite of being a large metropolitan city and being developed as a smart city, Chennai faces huge problems in terms of water resource availability. In 2019, Chennai faced an acute water shortage. On 19 June 2019, Chennai city officials declared that "Day Zero", or the day when almost no water is left, had been reached, as all the four main reservoirs supplying water to the city had run dry.

OBJECTIVE

The objective of the project is to visualize and analyse available historical data about water resources and water management in Chennai, Tamil Nadu.

The main objectives of this project are as follows:

1. Analyse individual water resources for Chennai.
2. Analyse water requirements of Chennai city through analysis of historical data about water resources for Chennai.
3. Perform data visualization over data to uncover trends and patterns in rainfall and water usage. For example, charts such as line charts will be plotted to visualize and analyse time-based data.
4. Visualize the water requirements of the city over the years. Various charts, such as bar charts and line charts, will be used to visualize the water usage of the city over the years.
5. Analyse the water crisis faced by Chennai city in 2019. Based on the various charts plotted using available data, we will try to find out

important patterns and trends that led to the crisis of 2019, by comparing it with previous years.

6. Perform statistical modelling and forecasting to predict water requirements in the future. We will use an ARIMA model to analyse time-based data. This will be important for forecasting the water requirements of the city in the near future.
7. Use time-series analysis algorithms to perform analysis. Time series analysis will be used for forecasting purposes. Time-series analysis models are widely used for non-stationary data, such as stock prices, weather conditions as well as economic parameters such as unemployment or GDP.

With the help of visualizations and statistical modelling techniques we will be able to predict whether the currently available water resources will be able to meet the water requirements of the residents of Chennai and surrounding areas.

PROBLEM STATEMENT

Chennai, the 6th most populous city of India, has a population of more than 1 crore residents as per the 2019 Census. It is one of India's most important metropolitan cities, and it's imperative that the city manages its water resources in a sustainable way to provide the residents with their daily needs. During 2019, Chennai faced a massive water crisis during which the situation became so worse that the government officials had to declare that 'Day Zero' has been reached owing to the depletion of all the city's four major water reservoirs.

The biggest reasons for this water scarcity were weak monsoons in the past few years, mismanagement of water resources by the city officials, use of deep borewells by the well-off sections of the populations.

The entire catastrophe, though resolved, left a huge question in its wake regarding the possibility of another such water scarcity in the city in the future. The idea that such a situation could arise again in the future invites the need of studies on the same matter, so that the city can be better equipped to deal with it beforehand itself.

Our project incepts from this case and would focus on visualizing the data relevant to it, culminating with the predictions for the likelihood of it happening again, along with some ways we can be prepared.

FUNCTIONAL REQUIREMENTS

1. HARDWARE REQUIREMENTS

a. Minimum of 4GB RAM will be required to efficiently perform analysis over the dataset and perform statistical modelling.
b. An i3 or higher processor is required to perform time-series analysis on the datasets.
c. Enough storage space (minimum of 1GB) is required to store datasets, visualizations and reporting dashboards for the project.

2. SOFTWARE REQUIREMENTS

a. **Anaconda:** The Anaconda navigator contains Spyder, which is an integrated development environment (IDE) for Python. We will be using this to perform data analysis and visualization using Python.
b. **Python 3.7.4**: Python 3.7.4 is a stable release of Python which we will be using for our analysis. It consists of all the required libraries that we will need for analysis.
c. **Matplotlib**: Matplotlib is a data visualization library for Python. We will use matplotlib to plot the various charts and graph required for analysis.
d. **Tableau Desktop:** We will be using Tableau Desktop, a data visualization and business intelligence tool, to make a data visualization dashboard for review presentations.
e. **Tableau Prep:** Tableau Prep will be used along with Tableau Desktop to prepare the dataset for data visualization.

DATA ABSTRACTION

The dataset used for this project is available at the link given below:

https://www.kaggle.com/sudalairajkumar/chennai-water-management

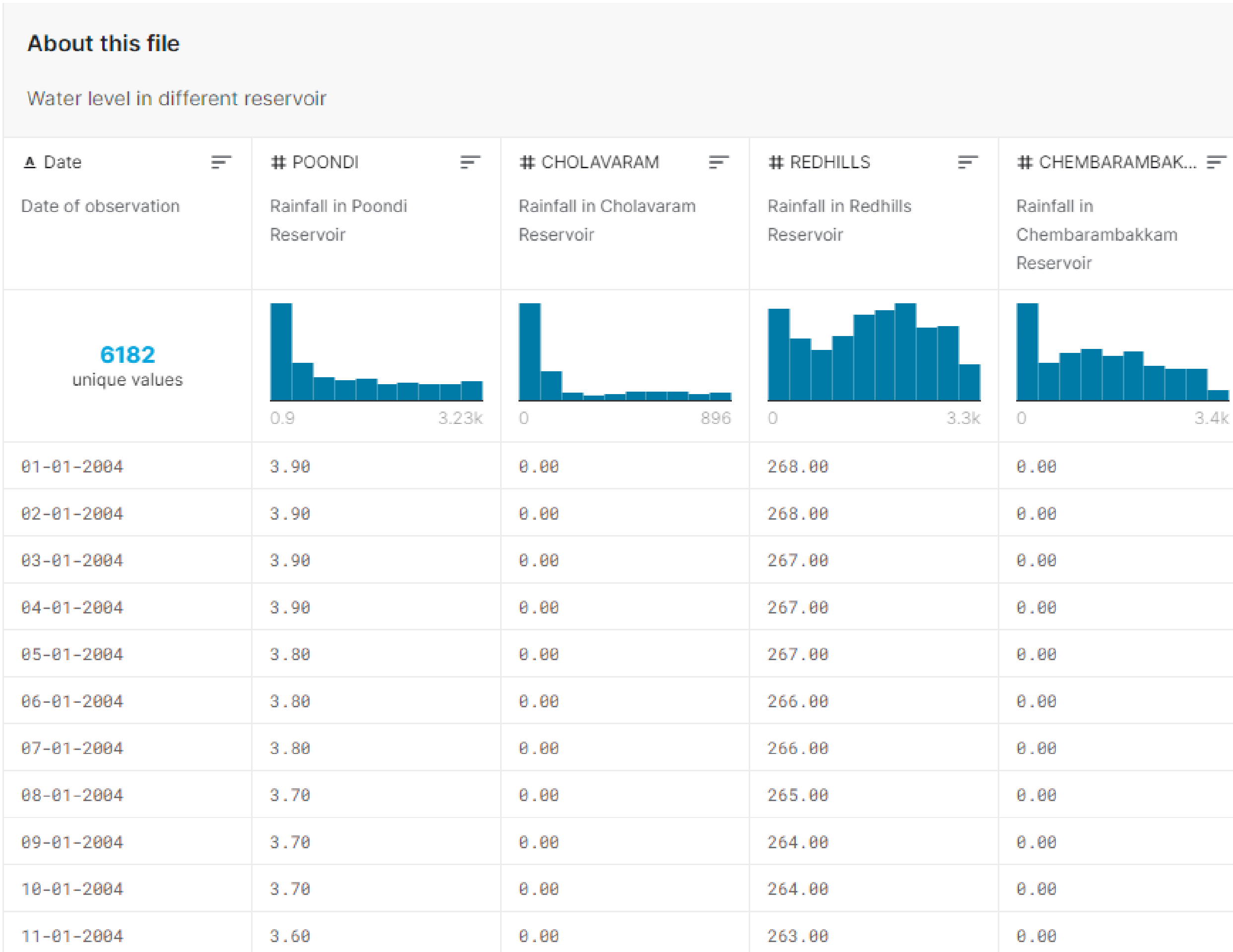
This dataset has details about the water availability in the four main reservoirs over the last 15 years

- Poondi Lake
- Cholavaram Aeri
- Red Hills
- Chembarambakkam Lake

The data about water availability is available on a daily basis. The dataset also includes data about the rainfall levels for Chennai over the last 15 years. This data will be used to perform required forecasting and analysis.

Water availability in 4 main reservoirs:

As we can see, this dataset about the volume of water available in the four main reservoirs near Chennai consists of daily data about water levels for all 4 reservoirs for the past 15 years (2004 - 2019). The dataset consists of 5836 unique data records.



About this file

Water level in different reservoir

Date	# POONDI	# CHOLAVARAM	# REDHILLS	# CHEMBARAMBAK...
Date of observation	Rainfall in Poondi Reservoir	Rainfall in Cholavaram Reservoir	Rainfall in Redhills Reservoir	Rainfall in Chembarambakkam Reservoir
6182 unique values	0.9 – 3.23k	0 – 896	0 – 3.3k	0 – 3.4k
01-01-2004	3.90	0.00	268.00	0.00
02-01-2004	3.90	0.00	268.00	0.00
03-01-2004	3.90	0.00	267.00	0.00
04-01-2004	3.90	0.00	267.00	0.00
05-01-2004	3.80	0.00	267.00	0.00
06-01-2004	3.80	0.00	266.00	0.00
07-01-2004	3.80	0.00	266.00	0.00
08-01-2004	3.70	0.00	265.00	0.00
09-01-2004	3.70	0.00	264.00	0.00
10-01-2004	3.70	0.00	264.00	0.00
11-01-2004	3.60	0.00	263.00	0.00

Rainfall in 4 main reservoirs:

The second dataset about the daily rainfall levels for the four major reservoirs is described above. It contains data about the daily amount of rainfall over the 4 reservoirs that are used to obtain water for Chennai city.

This dataset also has 5836 records of time-based data.

About this file

Rainfall level in different reservoirs

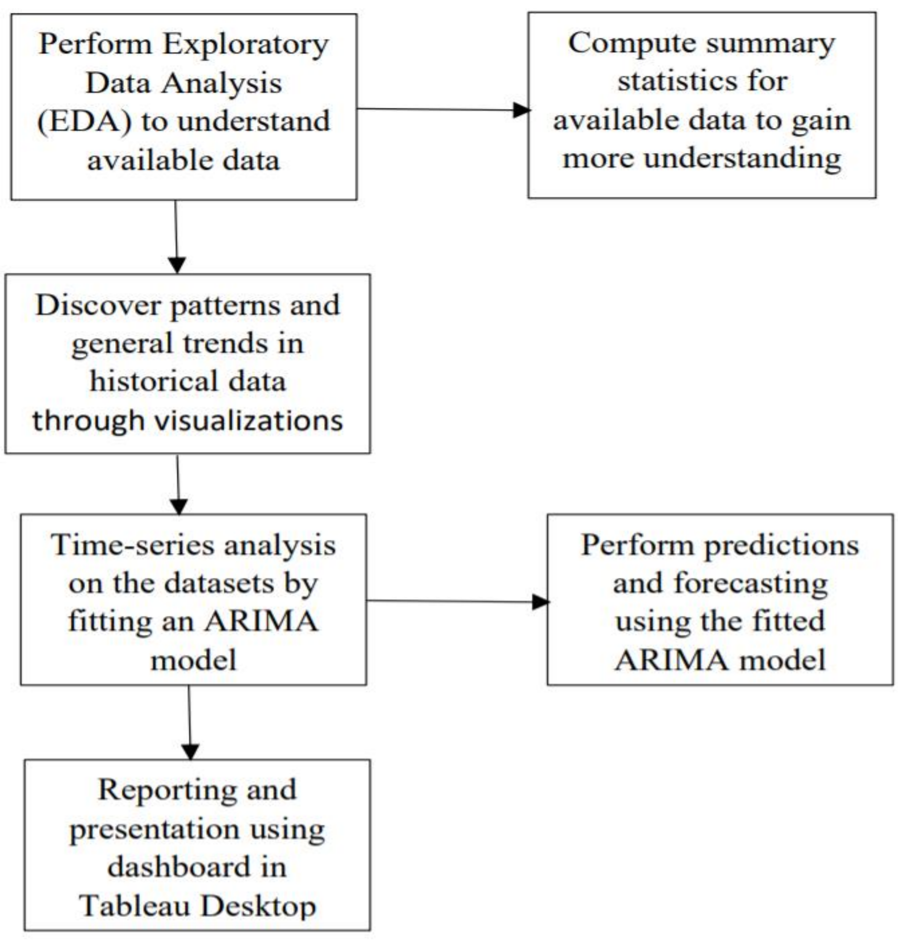
Date	POONDI	CHOLAVARAM	REDHILLS	CHEMBARAMBAK...
Date of observation	Rainfall in Poondi Reservoir	Rainfall in Cholavaram Reservoir	Rainfall in Redhills Reservoir	Rainfall in Chembarambakkam Reservoir
6182 unique values	0 – 300	0 – 293	0 – 320	0 – 475
01-01-2004	0	0	0	0
02-01-2004	0	0	0	0
03-01-2004	0	0	0	0
04-01-2004	0	0	0	0
05-01-2004	0	0	0	0
06-01-2004	0	0	0	0
07-01-2004	0	0	0	0
08-01-2004	0	0	0	0
09-01-2004	0	0	0	0
10-01-2004	0	0	0	0
11-01-2004	0	0	0	0

Both the datasets above will be used to perform analysis of water resource management and requirements for Chennai city. Time series forecasting can be applied since the data available to us is time-based.

DESIGN OF THE PROPOSED SYSTEM

In this project, we will be working with historical data about the Chennai water reservoir levels. We will make use of algorithms to predict whether the available water resources will be able to meet the water needs of Chennai and nearby people till next monsoon or not, starting with simple algorithms like time series forecasting.

CONCEPTUAL MODEL



ALGORITHM DESIGN

In this project, we have implemented time-series analysis of reservoir water level and rainfall data using an ARIMA model.

Time series analysis and forecasting is the use of a model to predict future values based on previously observed values. It comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data.

In time series forecasting, data is stored and recorded at a specific interval of time and then that data is analysed to forecast the future. It models the dataset based on the following components of time-based data:

1. Trend
2. Seasonality
3. Cyclicity
4. Irregularity

In order to perform statistical modelling on the datasets to perform time-series analysis, the algorithm that we have implemented can be described as follows:

- Add column ‘Total’ to reservoir water level dataset for total daily water level in all four reservoirs.
- Remove columns containing daily water level for individual reservoirs.
- Sort the records of the dataset by the ‘Date’ column.
- After sorting all records, set the ‘Date’ column as index for the data frame.
- Fit an ARIMA model for the reservoir water level dataset; o Select parameters for the ARIMA model as:
  - Seasonality (p) = 1
  - Trend (d) = 0
  - Noise (q) = 0

The required ARIMA model is available in the Python package statsmodels.api.

- Initialize ARIMA model with appropriate parameters and fit ARIMA model for the reservoir water level dataset.
- Plot diagnostics for the fitted ARIMA model to examine the accuracy.

Hence, we have fitted an ARIMA model to analyse the data based on time-series analysis. We have also performed forecasting using the fitted ARIMA model to forecast the total reservoir water level and total rainfall level.

TASK ABSTRACTION

MODULE DIAGRAM

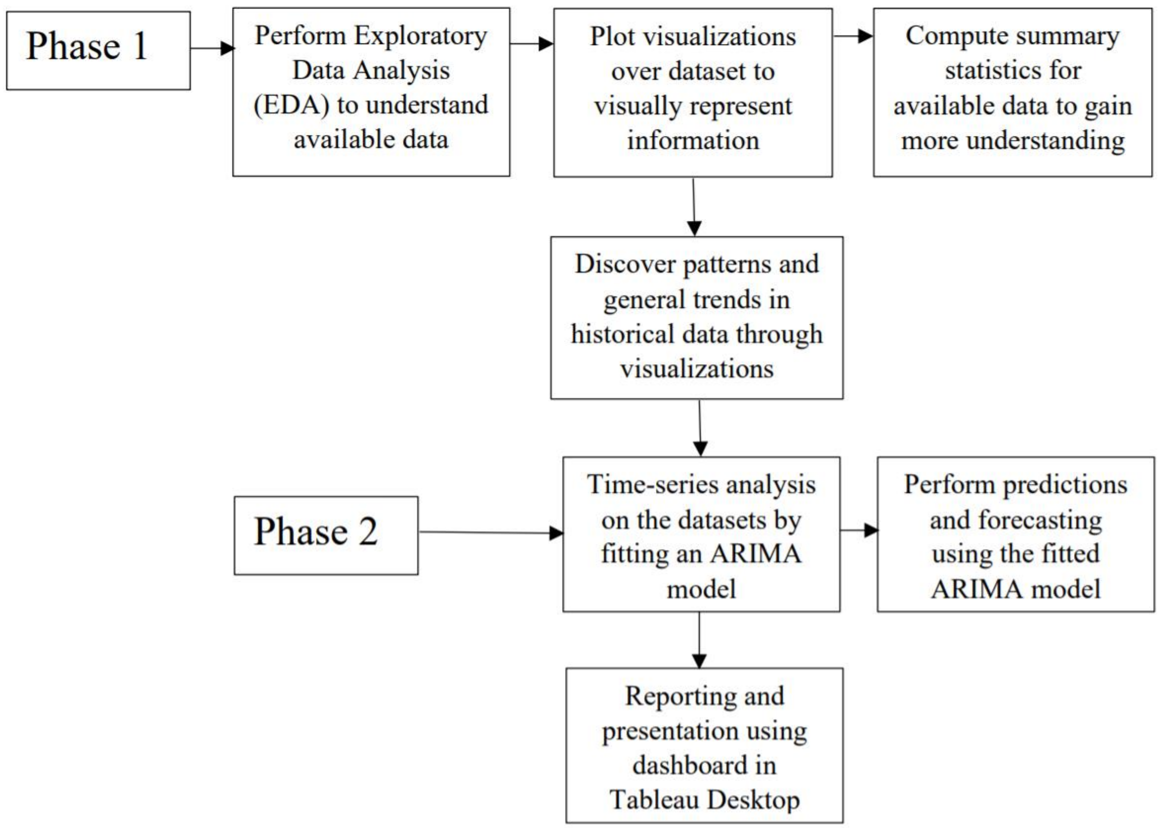
The implementation of this project has been divided into the following 2 phases:

1. Exploratory Data Analysis (EDA): In this phase of the project we have performed preliminary exploration of the datasets used in the project. Through this preliminary exploration, we gained a better understanding of the available data.

   We performed exploratory data analysis of the project with the help of 2 main techniques – summary statistics and visualizations. After gaining a better understanding of the available datasets, we started the next phase of the project which is described next.

2. Time-series Analysis: during this phase, we performed time-series analysis and forecasting for reservoir water level and rainfall datasets. We fitted and ARIMA model for the time-based data and performed forecasting with the fitted ARIMA model. Finally, we have consolidated all results of the project into a dashboard

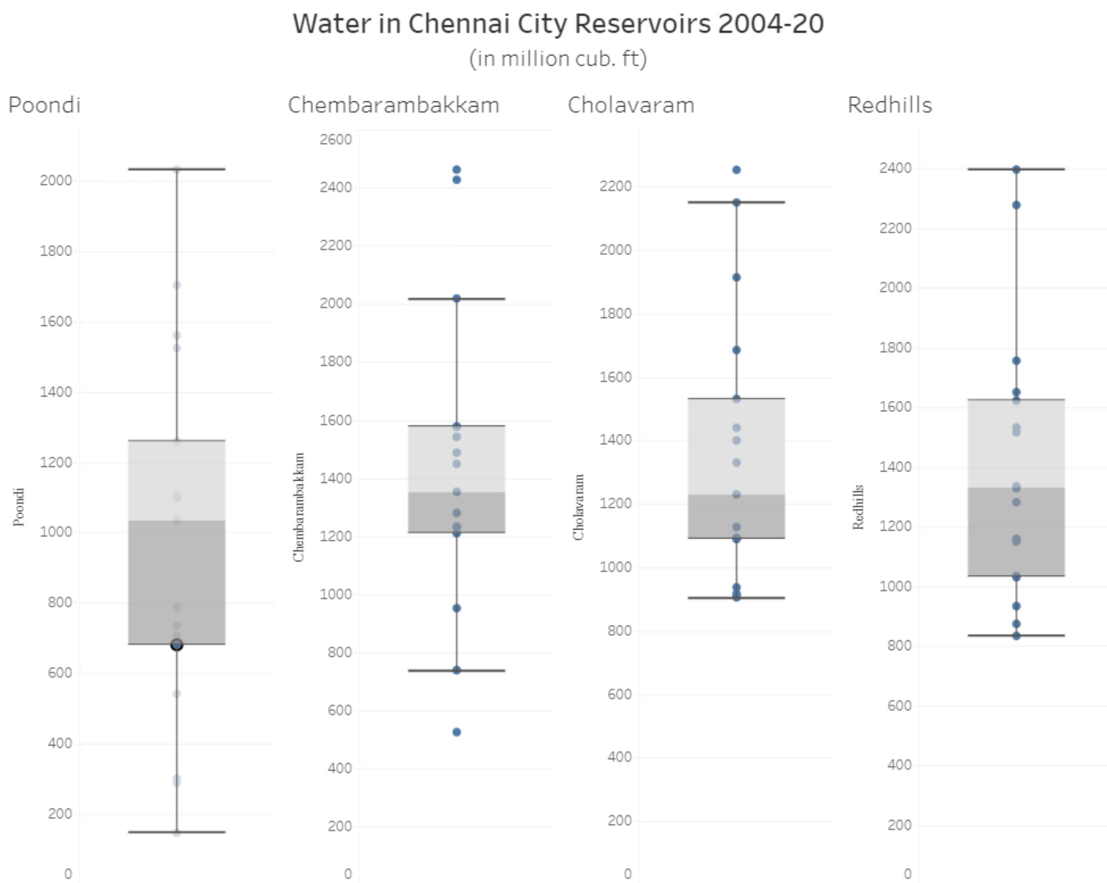
designed using Tableau Desktop for final presentation and reporting.



DASHBOARD IMPLEMENTATION

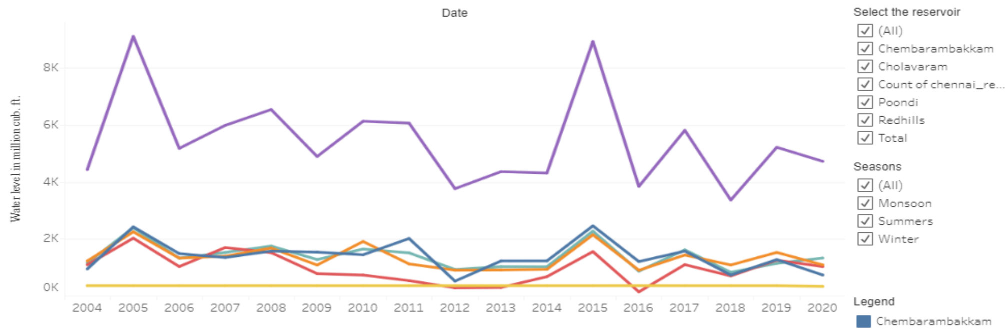
Dashboard created with Tableau Desktop

1)

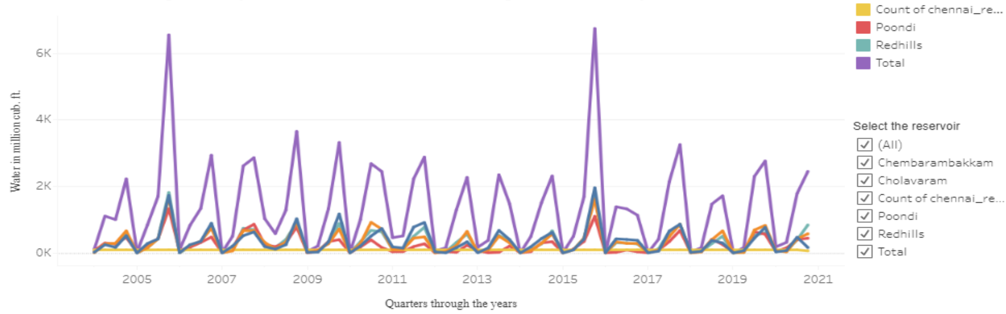


2)

Visualizing water levels over the years



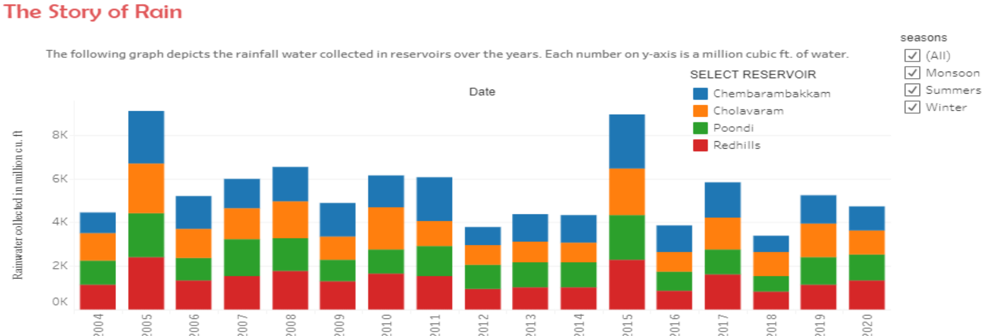
Let us introduce some granularity into the visualisation. The x-axis is granulated from a year to a Quarter.



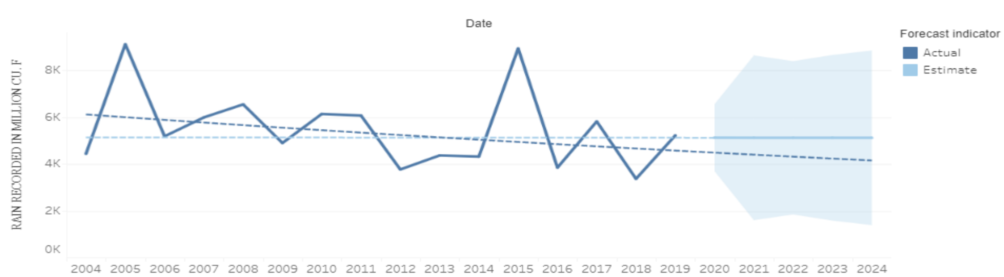
3)

The Story of Rain

The following graph depicts the rainfall water collected in reservoirs over the years. Each number on y-axis is a million cubic ft. of water.



So, what does mathematics have to say about the rainfall patterns in the future? The light blue region in the next plot depicts the possibility of error while the dotted lines show the trend of data.



RESULT ANALYSIS

Using the visualizations above, we can conclude the following about the reservoir water level and rainfall level datasets:

• Water level data has a seasonality that can be seen in the plots.

• Although this trend has been decreasing, in 2019 only Poondi reservoir had some water left. All other 3 reservoirs were running dry in 2019, during the water crisis in Chennai.

• During the month of December 2014, we can observe a sudden growth in water level. This is unnatural and the water must have been supplied to Chennai city from some other region.

• Downfall in water level is significant during the months from February to July in 2019.

• Evidence of water shortage can be noticed in all 4 reservoirs during the years 2017 as well as 2018.

• The dry period generally starts in the month of October and lasts until the month of December.

• Total water availability is very low in 2019. Periodic cycle is disturbed in this phase.

• A significant downfall in total water level has been observed since 2015 and keeps decreasing till 2019.

• Signs of heavy correlation are observed in the correlation matrices for reservoir water level and rainfall level. Seasonality can be explained using this evidence.

• The forecasted values for reservoir rainfall level indicate that during the next few years the rainfall is expected to be average.

• Also, forecasted values for reservoir water level show that the total water level across all reservoirs is expected to be lower than the historical average.

• These observations indicate the likelihood of a water crisis occurring in Chennai. Therefore, it may be required that water is conserved to deal with a likely crisis situation.

Hence, through our exploratory data analysis and time-series analysis, we can conclude that there is a seasonality and general trend in water and rainfall level for reservoirs. With the help of forecasts, we can say that water availability may be reduced due to average rainfall and below average reservoir water levels in the coming few years. Therefore, authorities and citizens should prepare adequately to deal with a probable water crisis in the near future.

CONCLUSION

With this project, we have successfully analyzed the historical data for water resources available in Chennai and discovered important patterns and trends in the data. Various statistical modelling techniques, including time series analysis and forecasting, have been used as part of the data analysis. The results of the project and data analysis can prove to be useful in many ways.

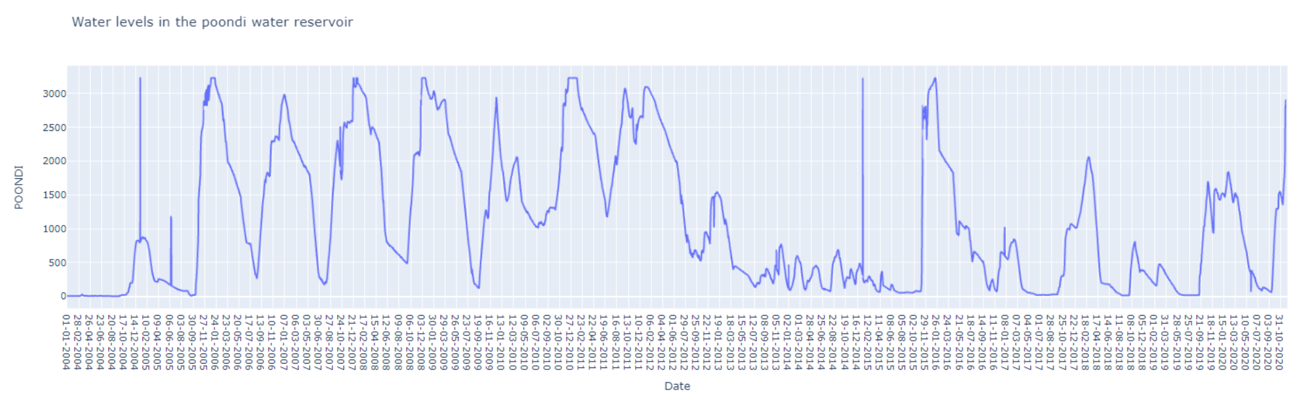
APPENDIX

- **SCREENSHOTS**

1) Exploratory Data Analysis

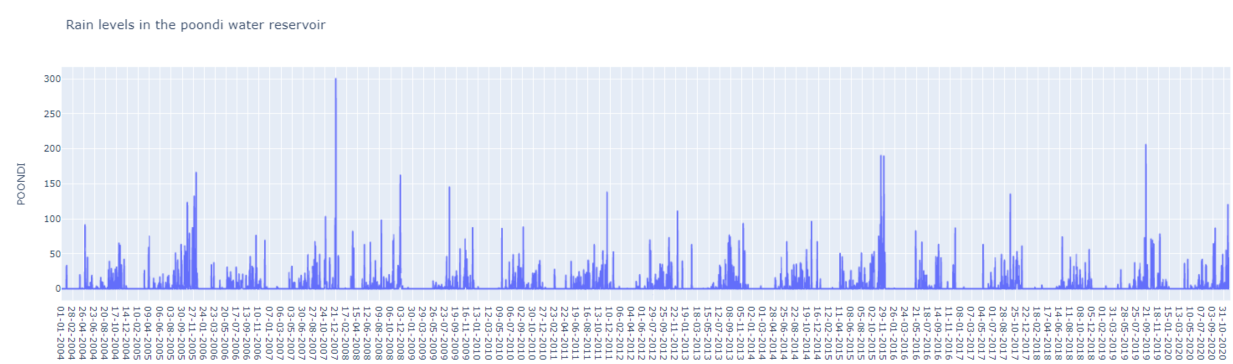
a) Poondi Water Levels

```
fig = px.line(lvl, x = "Date", y= "POONDI", title = "Water levels in the poondi water reservoir")
fig.show()
```



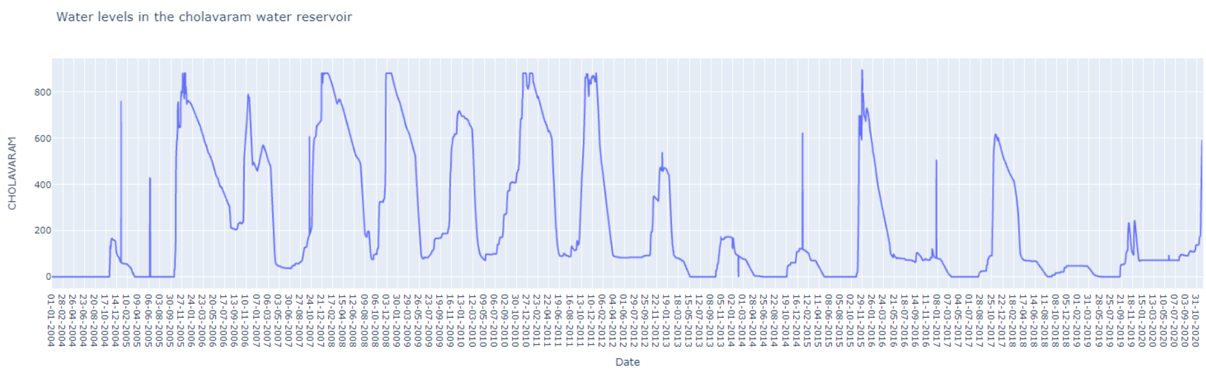
b) Poondi Rainfall Levels

```
fig = px.line(rain, x = "Date", y= "POONDI", title = "Rain levels in the poondi water reservoir")
fig.show()
```



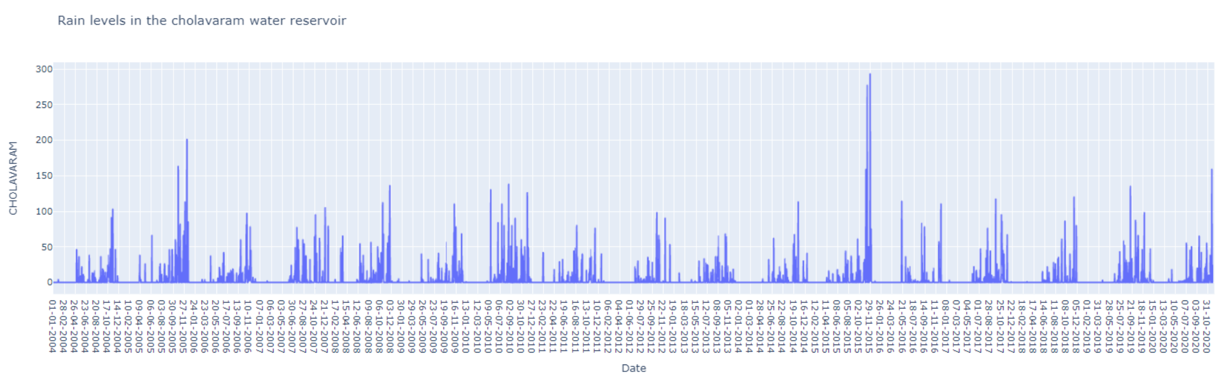
c) Cholavaram Water Levels

```
fig = px.line(lvl, x = "Date", y= "CHOLAVARAM", title = "Water levels in the cholavaram water reservoir")
fig.show()
```



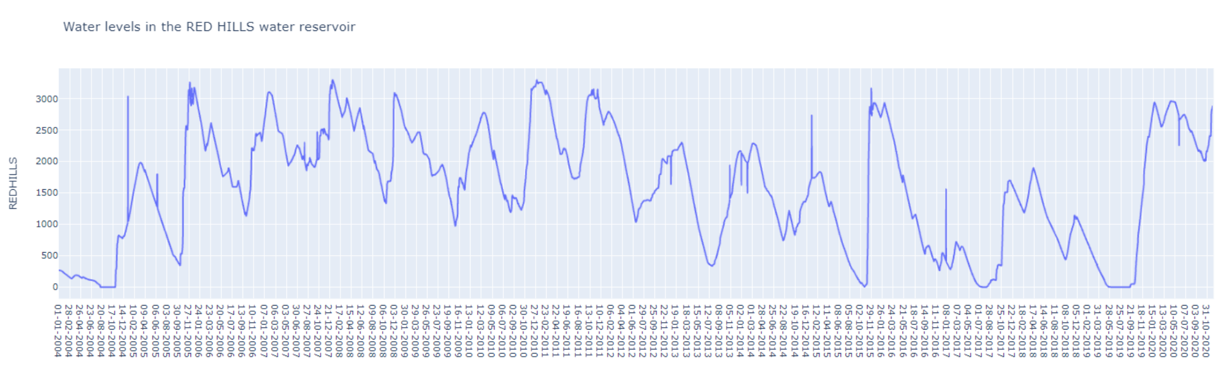
d) Cholavaram Rainfall Levels

```
fig = px.line(rain, x = "Date", y= "CHOLAVARAM", title = "Rain levels in the cholavaram water reservoir")
fig.show()
```



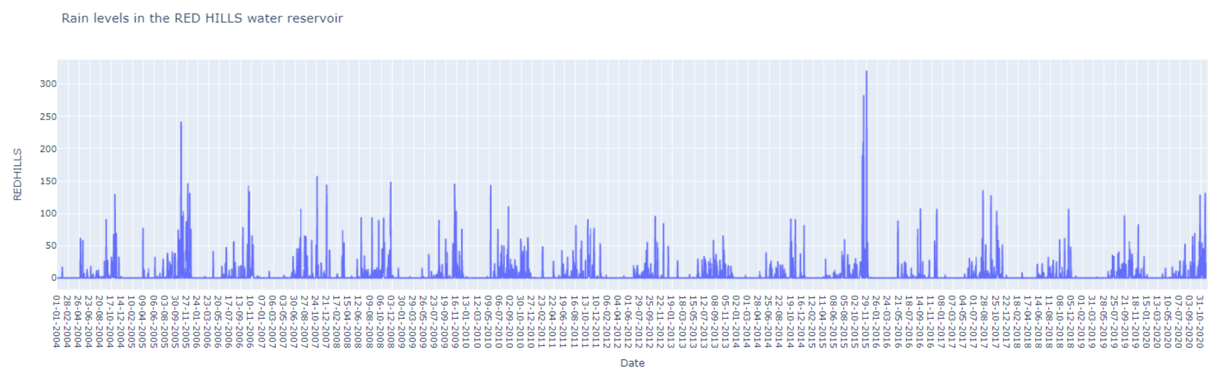
e) Red Hills Water Levels

```
fig = px.line(lvl, x = "Date", y= "REDHILLS", title = "Water levels in the RED HILLS water reservoir")
fig.show()
```



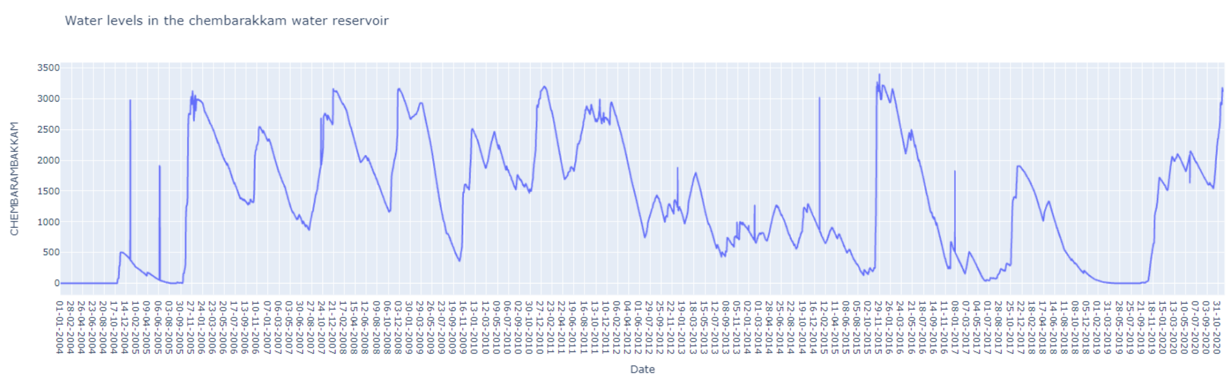
f) Red Hills Rainfall Levels

```
fig = px.line(rain, x = "Date", y= "REDHILLS", title = "Rain levels in the RED HILLS water reservoir")
fig.show()
```



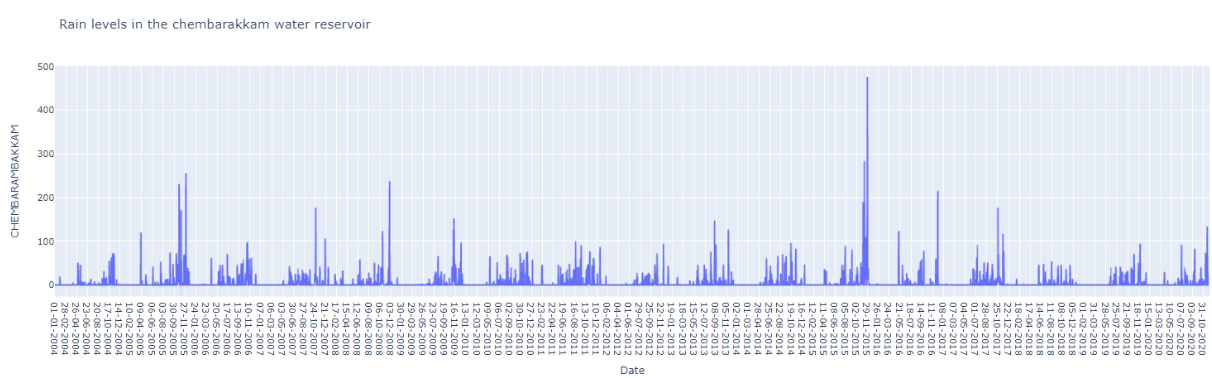
g) Chembarakkam Water Levels

```
fig = px.line(lvl, x = "Date", y= "CHEMBARAMBAKKAM", title = "Water levels in the chembarakkam water reservoir")
fig.show()
```



h) Chembarakkam Rainfall Levels

```
fig = px.line(rain, x = "Date", y= "CHEMBARAMBAKKAM", title = "Rain levels in the chembarakkam water reservoir")
fig.show()
```



2) Correlation

```
corr = lvl.corr()
corr.style.background_gradient(cmap='PuBu')
```

	POONDI	CHOLAVARAM	REDHILLS	CHEMBARAMBAKKAM
POONDI	1.000000	0.823638	0.754660	0.852809
CHOLAVARAM	0.823638	1.000000	0.690272	0.776602
REDHILLS	0.754660	0.690272	1.000000	0.862498
CHEMBARAMBAKKAM	0.852809	0.776602	0.862498	1.000000

```
#correlation between rainfall levels in all the reservoirs
corr = rain.corr()
corr.style.background_gradient(cmap='PuBu')
```

	POONDI	CHOLAVARAM	REDHILLS	CHEMBARAMBAKKAM
POONDI	1.000000	0.717860	0.706190	0.676785
CHOLAVARAM	0.717860	1.000000	0.900510	0.740898
REDHILLS	0.706190	0.900510	1.000000	0.774342
CHEMBARAMBAKKAM	0.676785	0.740898	0.774342	1.000000

- **SAMPLE CODING**

ARIMA Model

a) Data Cleaning

```
lvl = pd.read_csv("chennai_reservoir_levels.csv")
rain = pd.read_csv("chennai_reservoir_rainfall.csv")
```

```
#creating a new column of total in both datasets
lvl['Total'] = lvl.POONDI+lvl.CHOLAVARAM + lvl.REDHILLS + lvl.CHEMBARAMBAKKAM
rain['Total'] = rain.POONDI + rain.CHOLAVARAM + rain.REDHILLS + rain.CHEMBARAMBAKKAM
```

```
#removal of unneccessary columns
cols = ['POONDI','CHOLAVARAM', 'REDHILLS', 'CHEMBARAMBAKKAM']
lvl.drop(cols, axis = 1, inplace = True)
rain.drop(cols, axis = 1, inplace = True)
```

```
#sorting all datasets by date
lvl = lvl.sort_values('Date')
lvl = lvl.set_index('Date')

rain = rain.sort_values('Date')
rain = rain.set_index('Date')
```

b) Modelling ARIMA

```
#modelling the ARIMA
#ARIMA is used as ARIMA(p,q,d) where
#p, q, d are the seaonality of data, noise in data and trend in data

p = q = d = range(0,2)
pdq = list(itertools.product(p,d,q))
seasonal_pdq = [(x[0],x[1],[2],12) for x in list(itertools.product(p,d,q))]
```

```
print('Examples of parameter combinations for seasonal ARIMA as in this case...')
print('SARIMA: {} x {}'.format(pdq[1], seasonal_pdq[1]))
print('SARIMA: {} x {}'.format(pdq[1], seasonal_pdq[2]))
print('SARIMA: {} x {}'.format(pdq[2], seasonal_pdq[3]))
print('SARIMA: {} x {}'.format(pdq[2], seasonal_pdq[4]))
```

```
Examples of parameter combinations for seasonal ARIMA as in this case...
SARIMA: (0, 0, 1) x (0, 0, [2], 12)
SARIMA: (0, 0, 1) x (0, 1, [2], 12)
SARIMA: (0, 1, 0) x (0, 1, [2], 12)
SARIMA: (0, 1, 0) x (1, 0, [2], 12)
```

```
#fitting the ARIMA model and summarisation

lvl_model = sm.tsa.statespace.SARIMAX(lvl, order=(1,0,0), seasonal_order=(1,1,1,12), enforce_stationary=False, enforce_invertibility = False)
lvl_results = lvl_model.fit()
```

c) Prediction

Rainfall Levels

```
b=rain_results.predict(1,6800)
b.plot(figsize=(5,5))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f63396f6e10>
```