**Data Preprocessing**

Introduction

**Overview of the Datasets**

There are multiple datasets. One dataset contains value of weather parameters such as **temperature**, **humidity** etc. of one location over a period of 10 years . Another dataset contains values of pollution parameters such as **AQI, PM 2.5** and **PM10** over the last 10 years of the same location. Third dataset contains values of **congestion** **index** and special events occurring in the same location over a period of 10 years.

**Preprocessing Objectives**

The goal of preprocessing is to **clean, transform, and structure** the dataset for accurate visualization and analysis. The key objectives include:

* **Handling Missing or Inconsistent Data** – Removing or imputing missing values to ensure data integrity.
* **Standardizing Data Formats** – Converting timestamps and numerical values into a uniform format.
* **Filtering & Aggregating Data** – Extracting relevant records and summarizing values for meaningful insights.
* **Enhancing Data for Visualization** – Structuring the dataset to support effective charting and comparison.

By preprocessing the dataset efficiently, we aim to improve the reliability of the analysis and enhance the performance of real-time data visualization.

**Data Description**

The Dataset is extracted from Kaggle.in.

First dataset contains pollution data having fields like pm2.5 values , pm10 values , Windspeed ,Temperature and AQI values of a location in Delhi over a period of 10 years as shown below.

It contains the data for the following parameters from June 2016 to November 2023. The parameters included are : Timestamp, PM 2.5 value , PM 10 value , AQI , temperature , humidity and congestion index . It contains around 34,000 values for the location of Delhi .

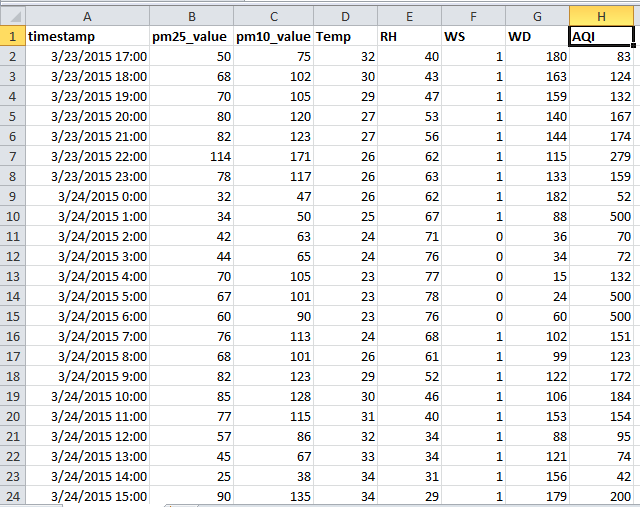


Fig 1.a.Screenshot of some data used

Second dataset contains values of weather parameters like temperature , humidity , wind speed etc. of same location in Delhi over the last 10 years

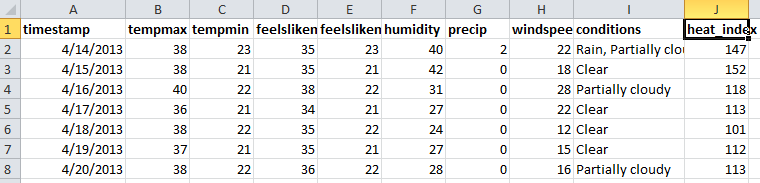


Fig 1.b . Sample of Weather data used

Third dataset contains values of traffic congestion index and special events of the same location in Delhi over a period of 10 years .

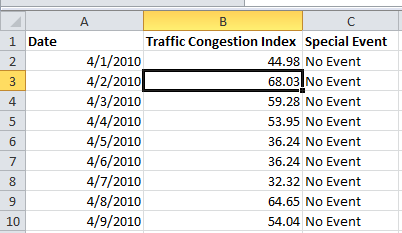


Fig 1.c . Sample of Traffic data used

**Data Cleaning**

The raw datasets was loaded into MYSQL database tables by the following load statement:

LOAD DATA INFILE 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/delhi\_traffic\_congestion.csv'

into table events\_data

fields terminated by ',' enclosed by '"'

lines terminated by '\n'

ignore 1 rows

(@date,@congestion\_index,@special\_event)

SET

date = STR\_TO\_DATE(@date,'%Y-%m-%d'),

congestion\_index = NULLIF(@congestion\_index,''),

special\_event = NULLIF(@special\_event,'');

The dataset originally had around 1lakh values . But more than 60% of the values had null values for one of the parameters . So the data was removed and trimed down to 34,000. To counter the problems of duplicates and to accumulate the values in a single place , a database was made and using the queries the removal of duplicates was done and the different datasets were merged and the data was put into a single csv file based on the timestamp by the following query :

SELECT 'timestamp','pm25\_value','pm10\_value','AQI','Temperature','tempmax','tempmin','feelslikemax','feelslikemin','humidity','precip','windspeed','conditions','heat\_index','congestion\_index','special\_event'

UNION ALL

select raw\_data.timestamp,pm25\_value,pm10\_value,AQI,temp,tempmax,tempmin,feelslikemax,feelslikemin,RH,precip,windspeed,conditions,heat\_index,congestion\_index,special\_event from raw\_data,weather\_data,events\_data where DATE\_FORMAT(raw\_data.timestamp, '%d %m %Y')= DATE\_FORMAT(weather\_data.timestamp, '%d %m %Y') and DATE\_FORMAT(raw\_data.timestamp, '%d %m %Y')= DATE\_FORMAT(events\_data.date, '%d %m %Y') and DATE\_FORMAt(raw\_data.timestamp,'%H:%i')='11:00'

INTO OUTFILE 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/11\_values.csv'

FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES TERMINATED BY '\n';

**Data Transformation**

In order to make sure data consistency occurs in the dataset , data has been sorted according to the values of the various parameters. Also repeating values of a parameter for different days have been normalized . A new parameter called special events had been added to see if an important event affects the pollution levels of the city . For better data visualization we have made graphs in which the x-axis contains the date and the y-axis contains the parameter to be measured . The AQI value, PM 2.5 value, PM 10 value and temperature have been taken in order to make the graph. Chart.js library have been used in order to make the graph.

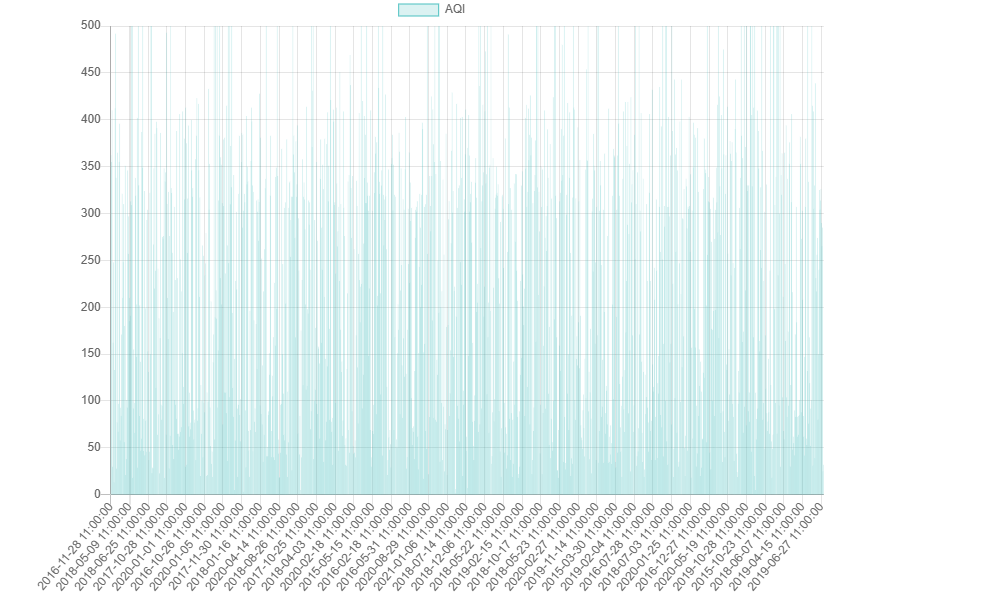


Fig 2.a . Graph for AQI vs Time

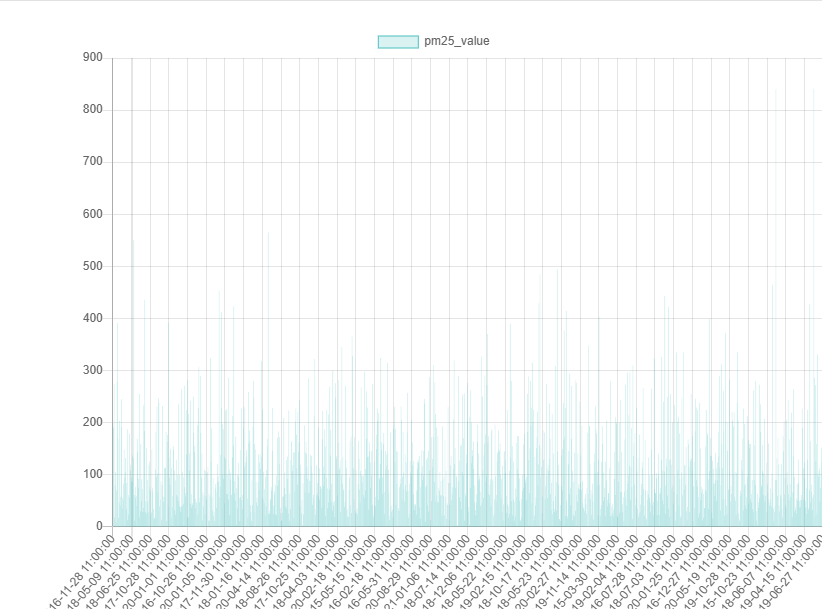


Fig 2.b. Graph for PM 2.5 value vs Time

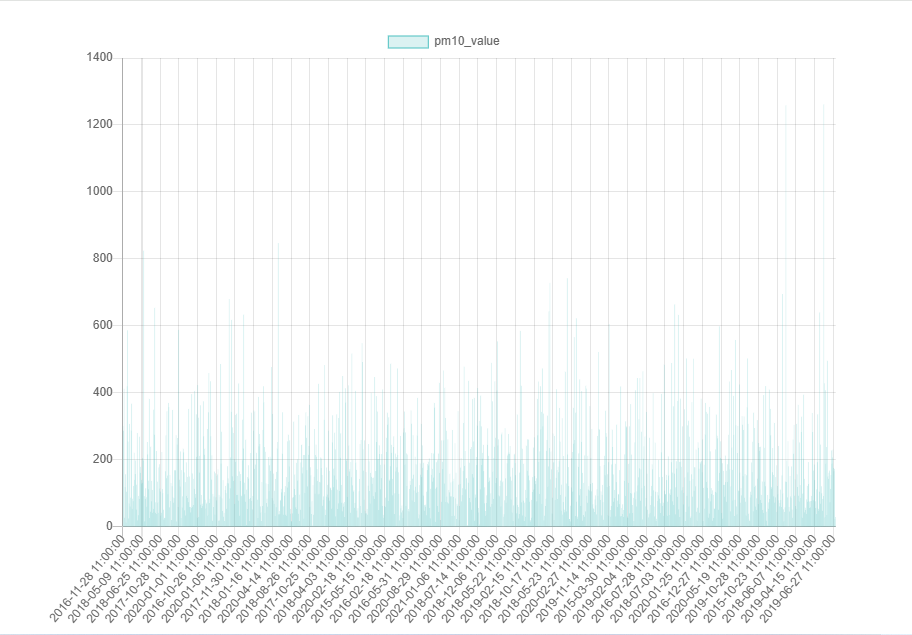


Fig 2.c. Graph for PM 10 value vs Time

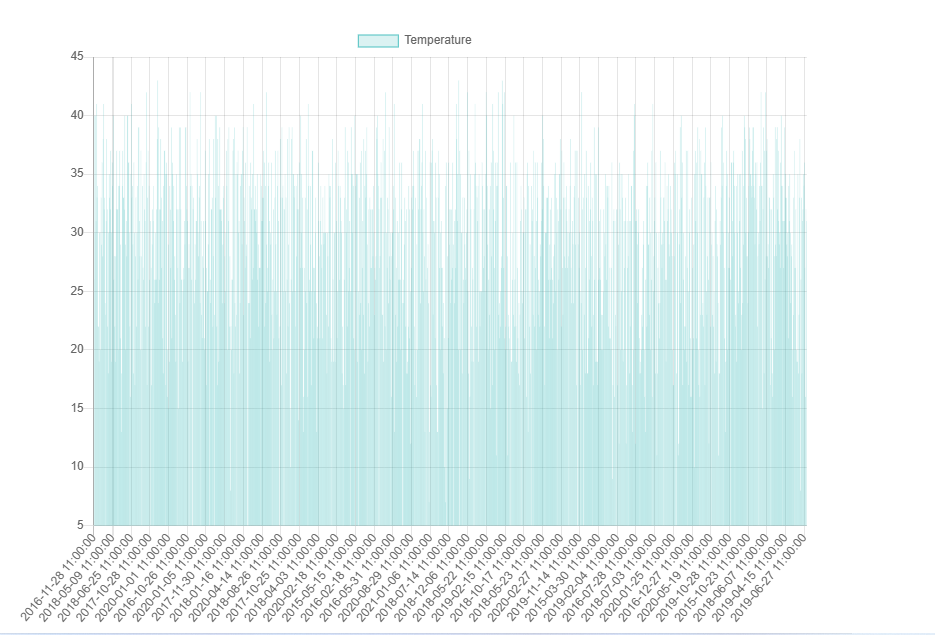


Fig 2.d . Graph for Temperature vs Time

**Data Integration and Challenges**

The Data was integrated by extracting different values from 3 datasets and then it is combined based on the dates of last 10 years in the 3 datasets .They have been combined using SQL queries and then normalized to form a single dataset by the following query :

SELECT 'timestamp','pm25\_value','pm10\_value','AQI','Temperature','tempmax','tempmin','feelslikemax','feelslikemin','humidity','precip','windspeed','conditions','heat\_index','congestion\_index','special\_event'

UNION ALL

select raw\_data.timestamp,pm25\_value,pm10\_value,AQI,temp,tempmax,tempmin,feelslikemax,feelslikemin,RH,precip,windspeed,conditions,heat\_index,congestion\_index,special\_event from raw\_data,weather\_data,events\_data where DATE\_FORMAT(raw\_data.timestamp, '%d %m %Y')= DATE\_FORMAT(weather\_data.timestamp, '%d %m %Y') and DATE\_FORMAT(raw\_data.timestamp, '%d %m %Y')= DATE\_FORMAT(events\_data.date, '%d %m %Y') and DATE\_FORMAt(raw\_data.timestamp,'%H:%i')='11:00'

INTO OUTFILE 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/11\_values.csv'

FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES TERMINATED BY '\n';

The challenges faced in this part was the huge amount of data present in the dataset. Also due to large values present in both the datasets , the cleaning process was difficult to execute . Also due to large amount of data , the other phases of data preprocessing were also affected and led to a higher complexity in making of the graphs.

Data had been normalized and cleaned using SQL methods and then converted into a single CSV file for better data interpretation and data visualization .

**Tools and Libraries used**

Excel – Used for making CSV files and basic data visualization

MYSQL database and SQL queries – Used for storing data and data preprocessing

Chart.js – Used in making graphs.

HTML and Javascript – Front-end and Back-end

**Challenges Faced**

1. **Missing or Incomplete Data**
   * **Issue**: Some rows might have missing values for AQI, PM2.5, PM10, or temperature.
   * **Solution**: Used interpolation, mean imputation, or removed rows with excessive missing values.
2. **Inconsistent Data Formatting**
   * **Issue**: Different date-time formats or inconsistencies in numerical values (e.g., commas instead of decimal points).
   * **Solution**: Standardized date formats and converted numerical values to a consistent format.
3. **Outliers in Data**
   * **Issue**: Extremely high or low values due to sensor errors or incorrect recordings.
   * **Solution**: Identified and filtered out outliers using statistical methods like Z-score or IQR.
4. **High Volume of Data**
   * **Issue**: Large dataset affecting performance in visualization.
   * **Solution**: Downsampled data by aggregating values on a daily/hourly basis.
5. **Encoding Issues in CSV**
   * **Issue**: Special characters or encoding mismatches causing read errors.
   * **Solution**: Ensured the file was encoded in UTF-8 and handled errors using encoding conversion.
6. **Handling Duplicates**
   * **Issue**: Duplicate rows affecting the accuracy of analysis.
   * **Solution**: Removed duplicate records based on timestamps and parameter values.

**Conclusion**

In this project, we focused on preprocessing the dataset to ensure data quality, consistency, and usability for visualization and analysis. The key preprocessing steps included handling missing values, standardizing data formats, removing outliers, optimizing data volume, and resolving encoding issues. These steps were crucial in ensuring the accuracy and reliability of the insights derived from the data.

By refining the dataset, we improved the performance of our visualization, enabling clearer trends and patterns to be identified. Proper preprocessing also reduced errors and inconsistencies, making the data suitable for further statistical analysis and machine learning applications.

Overall, these preprocessing techniques played a vital role in ensuring that the final dataset was clean, structured, and ready for effective decision-making and insights generation.