DAYANANDA SAGAR UNIVERSITY
SCHOOL OF ENGINEERING

Devarakaggalahalli, Harohalli, Kanakpura, Ramanagara Dt., Bangalore - 562112

REPORT FILE OF DATA ENGINEERING ASSIGNMENT - I

Air Quality Index

Submitted By

Khyathi M R ENG22CT0009

Under The Supervision Of

Prof. JUNAID MP

Asst. Professor, Department of CST, DSU

DEPARTMENT OF COMPUTER SCIENCE & TECHNOLOGY
SCHOOL OF ENGINEERING
DAYANANDA SAGAR UNIVERSITY
(2024-2025)

DAYANANDA SAGAR UNIVERSITY
SCHOOL OF ENGINEERING

Devarakaggalahalli, Harohalli, Kanakpura, Ramanagara Dt., Bangalore - 562112

CERTIFICATE

This is to certify that the work titled "Air Quality Index" is carried out by Khyathi M R-ENG22CT0009 Bonafide students of Bachelor of Technology in Computer Science and Technology at the School of Engineering, Dayananda Sagar University, Bangalore in partial fulfillment for the award of degree in Bachelor of Technology in Computer Science and Technology, during the year 2024-2025.

Junaid MP
Asst. Professor,
Dept. of CST,
School of Engineering,
Dayananda Sagar University.

Dr. M. Shahina Parveen
Chairperson, CST, Dept. of CST,
School of Engineering,
Dayananda Sagar University.

Date:

DAYANANDA SAGAR UNIVERSITY
SCHOOL OF ENGINEERING

Devarakaggalahalli, Harohalli, Kanakpura, Ramanagara Dt., Bangalore - 562112

DECLARATION

I, Khyathi M R-ENG22CT0009, a student of the sixth semester B.Tech in Computer Science and Technology, at School of Engineering, Dayananda Sagar University, hereby declare that the Data Engineering Assignment- I titled "Air Quality Index" has been carried out by me and submitted in partial fulfillment for the award of degree in Bachelor of Technology in Computer Science and Technology during the academic year 2024-2025.

Khyathi M R ENG22CT0009

Place:
Date:

ACKNOWLEDGEMENT

It is a great pleasure for us to acknowledge the assistance and support of many individuals who have been responsible for the successful completion of this DATA ENGINEERING ASSIGNMENT - 1. First, I take this opportunity to express my sincere gratitude to the School of Engineering Technology, Dayananda Sagar University for providing me with a great opportunity to pursue my bachelor's degree in this institution. I would like to thank **Dr. Udaya Kumar Reddy K R, Dean, School of Engineering Technology, Dayananda Sagar University** for providing this opportunity. It is an immense pleasure to express my sincere thanks to **Dr. M. Shahina Parveen, Chairperson, Department of Computer Science,and Technology, Dayananda Sagar University**, for providing the right academic guidance and motivating me during the course. I would like to thank my teacher **Prof. Junaid Mundichipparakkal, Asst. Professor, Department of Computer Science and Technology, Dayananda Sagar University**, for sparing his valuable time to extend help in every step of our course and the DATA ENGINEERING ASSIGNMENT – 1 , which paved the way for smooth progress and the fruitful culmination of the project.

ABSTRACT

The analysis of air quality plays a vital role in environmental monitoring and public health management, with accurate tracking of air quality indices (AQI) being essential for informed decision-making and risk assessment. This research focuses on analyzing a dataset collected from various air quality monitoring stations through real-time data extraction using API integration techniques.

The dataset encompasses crucial parameters such as location coordinates, pollutant concentrations (PM2.5, PM10, NO2, SO2, CO, and O3), timestamp, temperature, humidity, and wind speed. The study employs a structured data collection methodology, leveraging automated API requests to gather continuous air quality readings. The collected data undergoes rigorous cleaning, validation, and transformation to ensure accuracy and consistency. Outlier detection is performed for pollutant levels, while missing values are addressed using interpolation techniques. Standardization processes are applied to maintain uniformity in pollutant units, location names, and timestamp formats.

Key findings reveal significant air quality trends, including the influence of seasonal changes on pollutant levels, the impact of industrial zones and traffic congestion on AQI, and regional disparities in air quality patterns. The research also identifies potential inefficiencies in data reporting, such as gaps in monitoring and delayed updates from certain stations.

The insights derived from this study have practical applications in developing early warning systems, optimizing pollution control strategies, and informing policy decisions for environmental agencies. Additionally, this research contributes to predictive modeling for future air quality forecasting, supporting proactive measures for public health protection. The study emphasizes the need for standardized data reporting practices and enhanced digitalization in air quality monitoring to improve environmental management and mitigate health risks.

Keywords: Data Engineering, API Integration, Air Quality Index, Real-time Data, Environmental Monitoring

Contents

List of Figures

List of Tables

1 Introduction

- **Background:** The dataset is related to real-time air quality monitoring, collected from air quality stations across various locations in India, including states like Andhra Pradesh, Bihar, and Andaman and Nicobar Islands. It contains information about air pollutant levels such as PM10, PM2.5, NO2, NH3, CO, SO2, and OZONE. The data includes geographic details (latitude and longitude), city and station names, and pollutant concentration values (minimum, maximum, and average). This data is crucial for monitoring environmental health, identifying pollution sources, and supporting public health policies.

- **Problem Statement and Objectives:** The dataset aims to analyze air quality by assessing pollutant levels across regions. The goal is to identify pollution trends, understand geographic patterns, and support environmental and health-related decision-making.

  The key objectives are * Identifying common pollutants and their severity across cities.

  *Analyzing pollution trends based on geographic locations (states and cities).

  *Assessing the most affected areas and identifying locations with the highest pollutant levels

  - **Scope and Significance:** This dataset is valuable for environmental agencies, government bodies, and researchers to evaluate air quality and pollution hotspots. It helps in detecting areas with severe pollution, understanding temporal patterns, and guiding policy decisions on environmental protection.

- **Significance:**The dataset plays a crucial role in identifying and monitoring air quality, enabling environmental researchers and policymakers to enhance pollution control strategies. By analyzing pollutant distribution across cities and states, it aids in early detection of rising pollution levels, promoting timely intervention. Ultimately, it contributes to public health improvement and environmental conservation.

2 Dataset Overview

- **Description:** This dataset contains real-time air quality reports, including pollutant levels (PM10, PM2.5, NO2, NH3, CO, SO2, and OZONE), geographic details (country, state, city, station names, latitude, and longitude), and pollutant measurements (minimum, maximum, and average values).

- **Purpose of Collection:** The dataset is collected to monitor and analyze air quality across various regions in India. It helps assess pollution levels, identify the most affected areas.

- **Source of Data:** The data is sourced from data.gov.in , a public air quality monitoring stations across India, using an API.

- **Time Period & Frequency:** The data was collected on 24th February 2025, with a collection frequency of every 10 minutes. The real-time data capture provides insights into pollutant variations throughout the day.

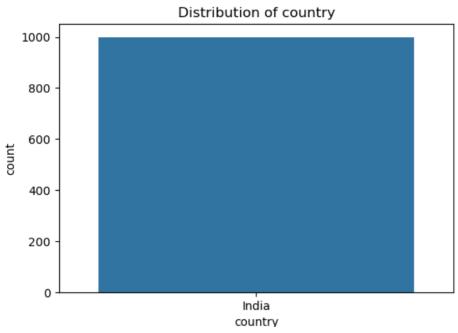
3 Dataset Structure and Features

- **Number of Rows:** 1000
- **Number of Columns:** 12
- **Data Storage Format:** CSV

4 Feature Explanations

4.1 Country

1. **Data Type:** String
2. **Statistical Data Type:** Nominal
3. **Range of Values:** Names of countries (e.g., "India")
4. **Distribution:**



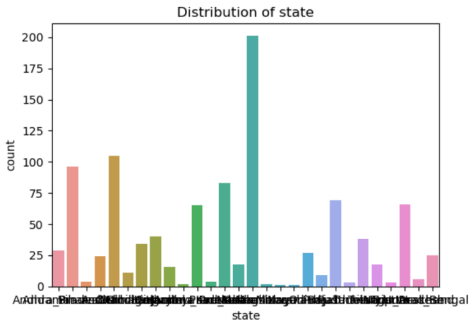
.

Missing Values: no missing values

4.2 State

1. **Data Type:** String
2. **Statistical Data Type:** Nominal
3. **Range of Values:** Names of states (e.g., "Andhra Pradesh," "Bihar," "Andaman and Nicobar Islands")

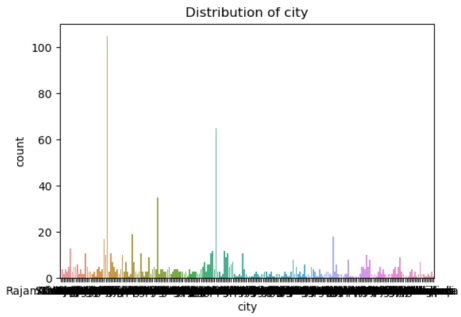
4. **Metric:** Frequency count of unique states
5. **Distribution:**



Missing Values: no missing values

4.3 City

1. **Data Type:** String
2. **Statistical Data Type:** Nominal
3. **Range of Values:** Names of cities (e.g., "Chittoor," "Kadapa," "Aurangabad")
4. **Metric:** Frequency count of unique cities
5. **Distribution:**

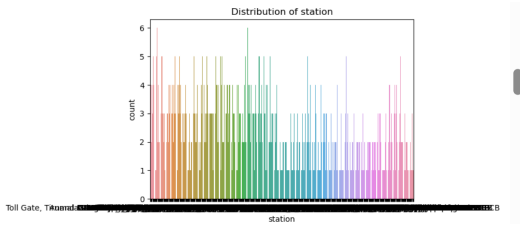


Missing Values: no missing values

4.4 Station

1. **Data Type:** String
2. **Statistical Data Type:** Nominal
3. **Range of Values:** Names of monitoring stations (e.g., "Gangineni C," "Yerramukka")

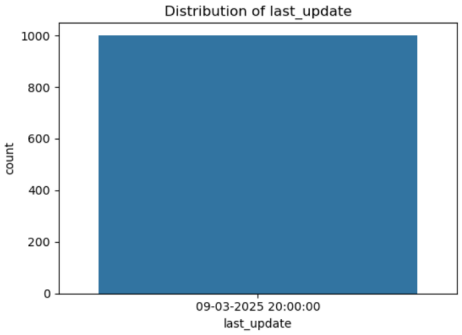
4. **Metric:** Count of unique drug reactions reported.
5. **Distribution:**



Missing Values: no missing values

4.5 Last update

1. **Data Type:** String
2. **Statistical Data Type:** Ordinal (timestamps in a non-uniform format)
3. **Range of Values:** The latest recorded time for each pollutant reading (e.g., "" placeholders in the current data, indicating missing or hidden values) .
4. **Metric:** Time intervals between updates
5. **Distribution:**



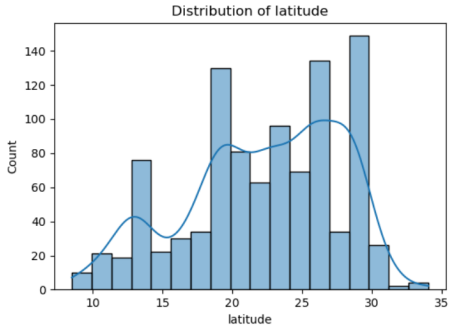
Missing Values: no missing values

4.6 Latitude

Provide an explanation here.

1. **Data Type:** Float
2. **Statistical Data Type:** Continuous

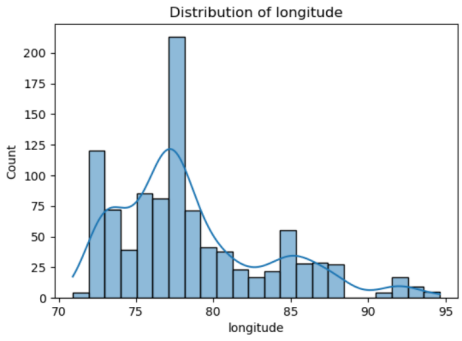
3. **Range of Values:**Approx. 10.0 to 30.0 (representing latitudes of Indian cities)
4. **Metric:** Mean and standard deviation can be used to understand the spatial spread of stations
5. **Distribution:**



Missing Values: no missing values

4.7 Longitude

1. **Data Type:** Float
2. **Statistical Data Type:** Conitnuous
3. **Range of Values:**70.0 to 90.0 (representing longitudes of Indian cities)
4. **Metric:**Used along with latitude for geographical mapping and spatial analysis.
5. **Distribution:**

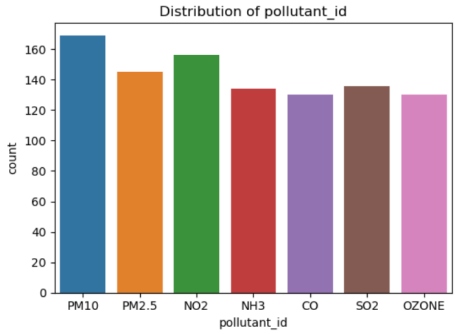


Missing Values: no missing values

4.8 Pollutant ID

Provide an explanation here.

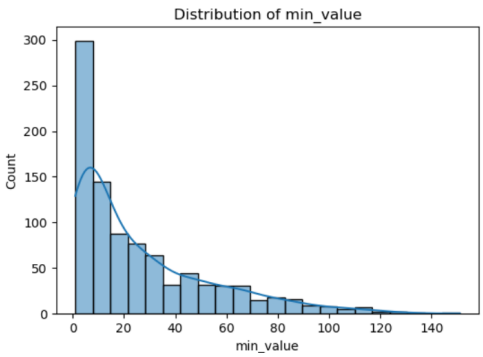
1. **Data Type:** String
2. **Statistical Data Type:** Categorical
3. **Range of Values:** Includes pollutants like PM10, PM2.5, NO2, NH3, CO, SO2, and OZONE
4. **Metric:**Count of each pollutant to identify the most frequently monitored pollutants.
5. **Distribution:**



Missing Values: no missing values

4.9 Min Value

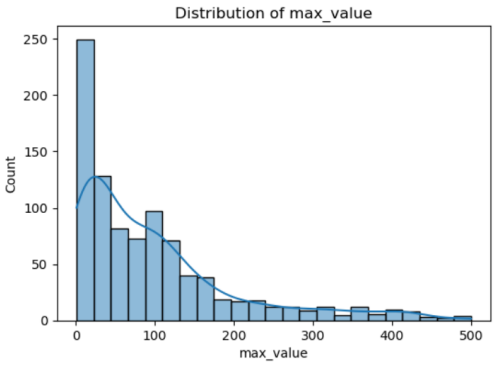
1. **Data Type:** Integer
2. **Statistical Data Type:** Continuous
3. **Range of Values:** 1 to 100 for valid pollutant readings
4. **Metric:** Used to track the lowest concentration recorded for each pollutant at a given time.
5. **Distribution:**



Missing Values: 73

4.10 Max Value

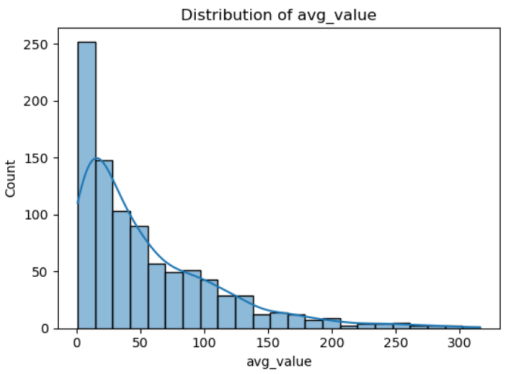
1. **Data Type:** Integer
2. **Statistical Data Type:** Continuous
3. **Range of Values:** 1 to 471 (based on pollutants like PM10)
4. **Metric:**Helps identify peak pollution levels and extreme environmental conditions.
5. **Distribution:**



Missing Values: 73

4.11 Avg Value

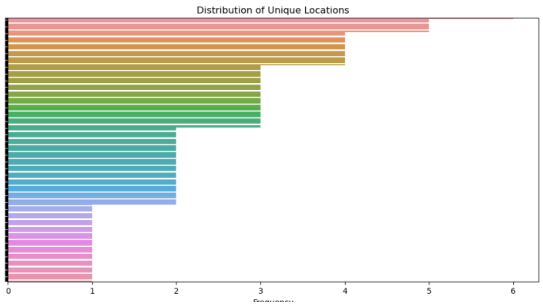
1. **Data Type:** Integer
2. **Statistical Data Type:** Continuous
3. **Range of Values:** 1 to 170 (pollutant averages over a time window)
4. **Metric:** Used for trend analysis, comparing pollution levels across regions.
5. **Distribution:**



Missing Values: 73

4.12 Location

1. **Data Type:** Tuple
2. **Statistical Data Type:** Nominal
3. **Range of Values:**Tuples representing (Longitude, Latitude) pairs for each monitoring station.
4. **Metric:** Count of occurrences of each unique location (latitude-longitude pair). Since Location is a tuple, it represents geographical coordinates, and the metric helps in identifying the distribution of locations across the dataset.
5. **Distribution:**



Missing Values: No missing values

5 Data Preprocessing and Cleaning

- **Handling Missing Values:**Many records in the dataset contain missing values, especially in fields like Min Value, Max Value, and Avg Value. For handling missing values, we have left entries marked as "NA" as-is, without imputation.
- **Outlier Detection and Treatment:**No explicit outlier detection has been implemented for pollutant values like PM10, PM2.5, NO2, etc.
- **Normalization and Encoding:** The dataset is fetched in a structured format (e.g., from an API) and converted into a Pandas DataFrame for analysis.No additional normalization or encoding has been applied to this dataset, but certain continuous columns (e.g., Min Value, Max Value, Avg Value) may require normalization if being used for machine learning models.
- **Feature Engineering:** Additional columns added: "Location".

6 Miscellaneous

- **Data Quality Issues:**Potential issues include inaccuracies in the reported pollution levels due to malfunctioning or miscalibrated monitoring stations. Additionally, discrepancies between

stations in different regions may lead to inconsistencies in air quality readings. Some pollutants may be underreported in certain regions due to the absence of monitoring stations. Missing data in certain pollutant values, marked as "NA", might affect completeness and reliability in specific analyses.

- **Limitations:** The dataset is limited to air quality data from 2024 and only includes stations located in Indian cities. It doesn't account for non-monitored regions or rural areas, potentially leading to incomplete representations of the overall air quality in India. Additionally, some stations may report incomplete data at certain times, as evidenced by the missing "NA" values for pollutants. Data depends on FDA reporting, which may not capture all real-world events.

- **Assumptions:**We assume the data collected by the monitoring stations is accurate and represents the conditions on the specific dates and times recorded. We also assume that the existing missing data (represented as "NA") can be handled with standard imputation techniques, or that it does not significantly affect the analysis.

- **Ethical Considerations:** The dataset does not contain any personally identifiable information (PII) such as names, addresses, or any other private individual data. The dataset only includes geographical data and air quality readings for public regions and monitoring stations. As the dataset is focused on environmental data, there are no significant privacy concerns.

  Public Domain: The data is publicly available from governmental and non-governmental monitoring agencies for air quality and is often shared under open-access policies. Proper attribution to the source of the data is encouraged in any publication or application utilizing this dataset.

References

1. https://link.springer.com/chapter/10.1007/978-981-10-5792-2$_8$

2.*https* : *//link.springer.com/article/*10.1007/*s*11157 − 010 − 9227 − 2

A Appendix

Make it empty if there is nothing here

A.1 Survey Report

This section has more details.

A.2 Interview Report

This section has more details.