Data Analaysis End To End Project Documentation

High Level Design (HLD)
(Air Pollution Data Analysis)

Document Version Control

Date Issued	Version	Description	Author
25-08-2022	HLD-V1.0	First Version of Complete HLD	Gokul

Revision Number: 1.0
Last Revision :25-08-2022

Contents:

Document Version Control .....
Abstract .....

1. Introduction .....
    - 1.1 Purpose of the Document
    - 1.2 Objective of HLD
    - 1.3 Scope of HLD

2. General Description .....
    - 2.1 Product Perspective & Problem Statement
    - 2.2 Data Requirements
    - 2.3 Tools Used
    - 2.4 Constrains

3. Design Details .....
    - 3.1 Process Flow
    - 3.2 Error Handeling / Exception Handling

4. Conclusion .....
5. References .....

Abstract:

India, a developing Nation that is home to 138 crore people, is among the world's urban agglomerations with the most toxic air. The magnitude of air pollution is massive. It causes devastating impacts on people's health, the city's environment, and economic well-being. Despite overwhelming evidence of the severity of air pollution and its consequences, however, India's policy measures remain weak. This paper identifies the most crucial gaps in policies and outlines a framework for creating more focused targets that will improve air quality in India.

Among all the cities in India, some of the worst levels of air pollution are seen in its capital territory, Delhi. he impacts are devastating, including in the degree of particulate matter concentrations in the air (environmental), reduction in life expectancy (health), and high costs that the state is incurring to resolve the crisis (economic). The main sources of air pollution in Delhi include vehicle exhaust, heavy industry such as power generation, small-scale industries like brick kilns, suspended dust on the roads due to vehicle movement and construction activities, open waste burning, combustion of fuels for cooking, lighting, and heating, and in-situ power generation via diesel generator sets. Compounding the problem are seasonal emissions from dust storms, forest fires, and open field fires during harvest season. Extreme air pollution from these sources affects millions of people in densely populated regions who are exposed to thick, toxic smog for long periods of time.

This study demonstrates the how air pollution in Delhi and other places of India affects the air quality makes life difficuilt to live and breath. Different analysis performed such as Exploratory Data Analysis and Descriptive Analysis on variety of matrices to get the key insights from this data based on which how serious we need to take air pollution and prevent our enviroinment.

1. Introduction:

This document will be used for documenting High-level designs of project.

1.1 Purpose of the Document

The purpose of this plan is to

- Describe different design app roaches.
- Describe different analysis approaches based on variety of matrics.
- Describe third party components/tools required for the system.
- Present complete Process Flow followed for this project.

1.2 Objective of HLD

1. To provide an overview of the entire system.
2. To provide introduction of Problem Perspective & Statement, Data Requirements, Tools used and many more.
3. To provide a module-wise breakup of the entire system.

1.3 Scope of HLD

This HLD covers all areas of the system.

2. General Description

2.1 Analysis Perspective & Problem Statement

India has been particularly vulnerable to air pollution over the last two decades, owing to population growth, increasing numbers of vehicles, use of fuels, inefficient transportation systems, poor land use patterns, industrialisation, and ineffective environmental regulations.
The objective of the project is to perform an exploratory data analysis, data pre-processing, & data cleaning and at the end, apply different Data Visualization techniques to get the meaningful insight from the given data. This project aims apply some amazing Python Libraries such as seaborn and matplotlib which will give a boost to our visual understanding of the data that's collected.

2.2 Data Requirements

Data Requirement completely depend on our problem.

In this project, to perform analysis, we are using the data that is been collected from https://data.gov.in/ which is the official website by the Indian government.
The features that are taken into consideration are:

Name	Description
State	Name of the states in India from were data is been updated
City	Respective Cities from the States of India
Last Update	The data on which the data is been updated
Pollution ID	The type of air pollutant emit in India
Pollution Avg	The Avg pollution rate on day to day basis

2.3 Tools Used.

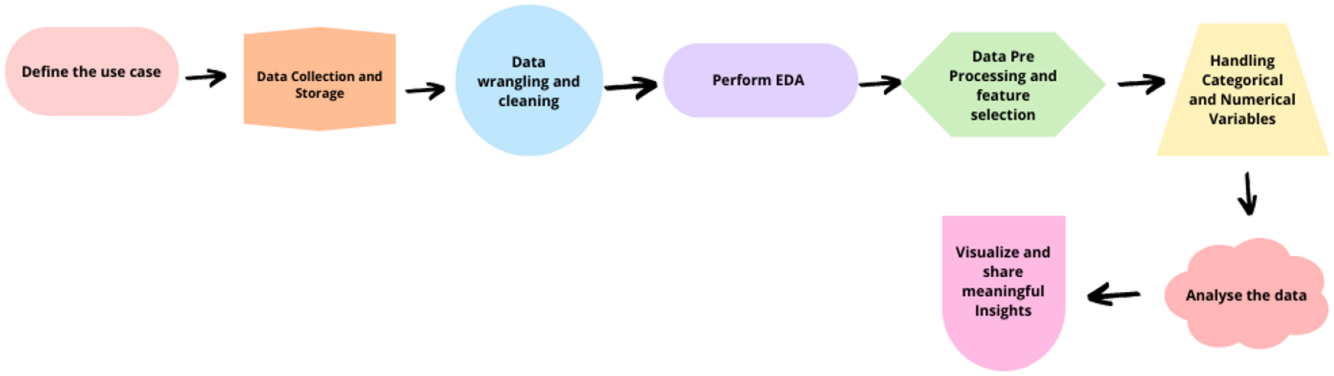
- **Website API** to fetch the data from website.
- **Google sheets** to temporarily store and append the data that's been collected every 5 hours.
- **AWS EC2** as a virtual machine to run the program every hour to collect the data.
- **AWD RDS** to store the collected Data.
- **Cronjob** as an schedules to run the py script every hour.
- **Git Bash** to intract with the EC2 ubuntu machine.
- **MySq**l workbench to query and keep a track of collected data.
- **Jupyter NoteBook** is used as an IDE.
- **Pandas and NumPy** are used for Data Manipulation & Pre-processing and Mathematical functions respectively.
- Exploratory data analysis is automated by **dataprep**.
- For visualization of the plots, Matplotlib, Seaborn, geopy and folium are used.
- **GitHub** is used as version control system.

2.3 Constrains.

The analysis must be user friendly, code must be neat & clean, EDA must be automated as much as possible because it will save huge amount of time. Moreover, users should not be required to have any of the coding knowledge as the insights they are looking for are mentioned in-detail with respective visuals.

3. Design Details

3.1 Process Flow.



3.1 Error Handleing / Exception Handling:

I have designed this project in such a way that, complete script is tested and runs multiple times to make sure that there is no error occurred during process flow. Additionally, we have also dismissed the un-necessary warnings to avoid confusion by using filterwarnings class from warnings module.

4. Conclusion

In this analytical project, we have examined a number of different matrices in order to assess the data we have gathered and make wise judgments that will benefit the environment and atmosphere, improve air quality, and generally make the world a better place to live. It has been discovered that -

When counting, it appears that ***Delhi has the highest number of pollutant emittance***; however, this may be an unfair comparison because, when considering the ***surface area of Delhi to the population***, it has the highest count.

Based on the data, it appears that each state contributes at a different level to different pollutant levels, which could be due to the geographical location of the states.

Let's take the example of Delhi itself, based on the data the pollutant emittent **'PM2.5'** and **'NO2'** have the highest count. It's understandable given the large number of vehicles and pollutants emitted by the industries themselves.

Infact, the air pollution in ***Delhi is so bad that, it has the total pollutant emit of Delhi is equal to rest of the six cities (Mumbai, Bengaluru, Kolkata, Hyderabad, Ghaziabad, Patna) combained***.

The fact that the pollution level rose from ***55 to over 80 in just three to five days is a striking indication that firecrackers can significantly raise the air pollution level***.

According to the above line map, the average ***air pollution level from November through February ranged from 50 to 70***.

But the months of ***November and December are when air pollution reaches its worst levels***. And during the months of January and February, the amount of air pollution typically drops from 50 to 60 to 40 to 30.

Addition to our above finds with the data we had I have found few interesting insights from the news articles -

- The restrictions on non-essential movement in the first few months of the COVID-19 pandemic led to a significant decline in air pollution levels across India.
- It helped achieve 95 percent of National Clean Air Program targets for 2024 in just 74 days in Delhi, Mumbai, Kolkata and Chennai, as emissions from the transport, construction and industrial sectors almost stopped and those from power plants reduced significantly.

- Air pollution, however, is not a one-time, short-term crisis; it is a recurring problem that requires long-term, holistic solutions.
- If the lockdown showed anything, it is that air pollution levels can be brought down dramatically if India focuses its energy towards a green recovery model that is less emissions-intensive.

5. References

1. Government website
2. Air Pollution artice
3. Delhi Air Pollution aritcle with respect to population.
4. Indian Air Pollution Wiki