Project Report on

Pollution Data Analysis

Submitted by:

Drishti

Abhishek Shukla

In partial fulfillment of completion of the course

Advanced Diploma in IT, Networking and Cloud Computing.

Under Guidance of:

Year 2022-2023

Abstract

The Pollution Data Analysis project focuses on leveraging data-driven methodologies to comprehensively analyze environmental pollution, with a primary emphasis on air and water quality. By employing statistical techniques, exploratory data analysis, time series analysis, and machine learning models, this project aims to unravel intricate patterns, trends, and contributing factors associated with pollution.

Acknowledgement

At this juncture of our journey, we wish to express our heartfelt gratitude to all those who have contributed to the creation and success of **"Pollution Data Analysis".** This project has been a labor of passion and dedication, and it would not have been possible without the unwavering support and guidance we have received.

First and foremost, we offer our thanks to the boundless creativity and inspiration that flows from the universe. We are grateful for the opportunity to embark on this venture.

We extend our sincerest appreciation to our mentors, **Mrs. Mala Mishra & Ms. Ankita Shukla**, whose wisdom and guidance have been instrumental in shaping the vision of **"Pollution Data Analysis".** Your support at every crucial turn has illuminated our path and fueled our determination to create a meaningful platform.

To our dedicated team of developers, designers, and content creators, we extend our deepest gratitude. Your tireless efforts, innovation, and creativity have breathed life into **"Pollution Data Analysis".** It is your collective dedication that has made this project a reality.

Our appreciation also goes to our colleagues and friends who provided invaluable insights and feedback during the development process. Your input has been instrumental in refining our ideas and enhancing the user experience.

We acknowledge the contributions of the broader IT community, whose open-source ethos has been a wellspring of knowledge and inspiration. The collaborative spirit of this community has been a guiding light.

Last but not least, we owe a debt of gratitude to our families and friends who have stood by us throughout this journey. Your unwavering support, encouragement, and belief in our vision have been our constant motivation.

ADVANCE DIPLOMA IN IT NETWORKING & CLOUD COMPUTING

The Advanced Diploma in IT Networking and Cloud Computing program offered by NSTI (W) Noida in collaboration with Edunet Foundation is a comprehensive course designed to equip students with advanced skills in information technology and cloud computing. This program covers a wide range of topics, including Computer Networking, Database Management, Virtualization, Cloud Technologies, and Cybersecurity. Students will gain hands-on experience through practical labs, workshops, and real-world projects, enabling them to excel in the rapidly evolving IT industry. Upon completion of the program, Graduates will have a strong foundation in both IT Fundamentals and Cloud Computing, making them highly sought-after professionals in the field.

Project Requirements

Project Name	**Pollution Data Analysis**
Languages Used	**Python**
Editor	**Jupyter Notebook, Google Colab**
Web Browser	**Google Chrome, Microsoft Edge**

Team Composition and Workload Division

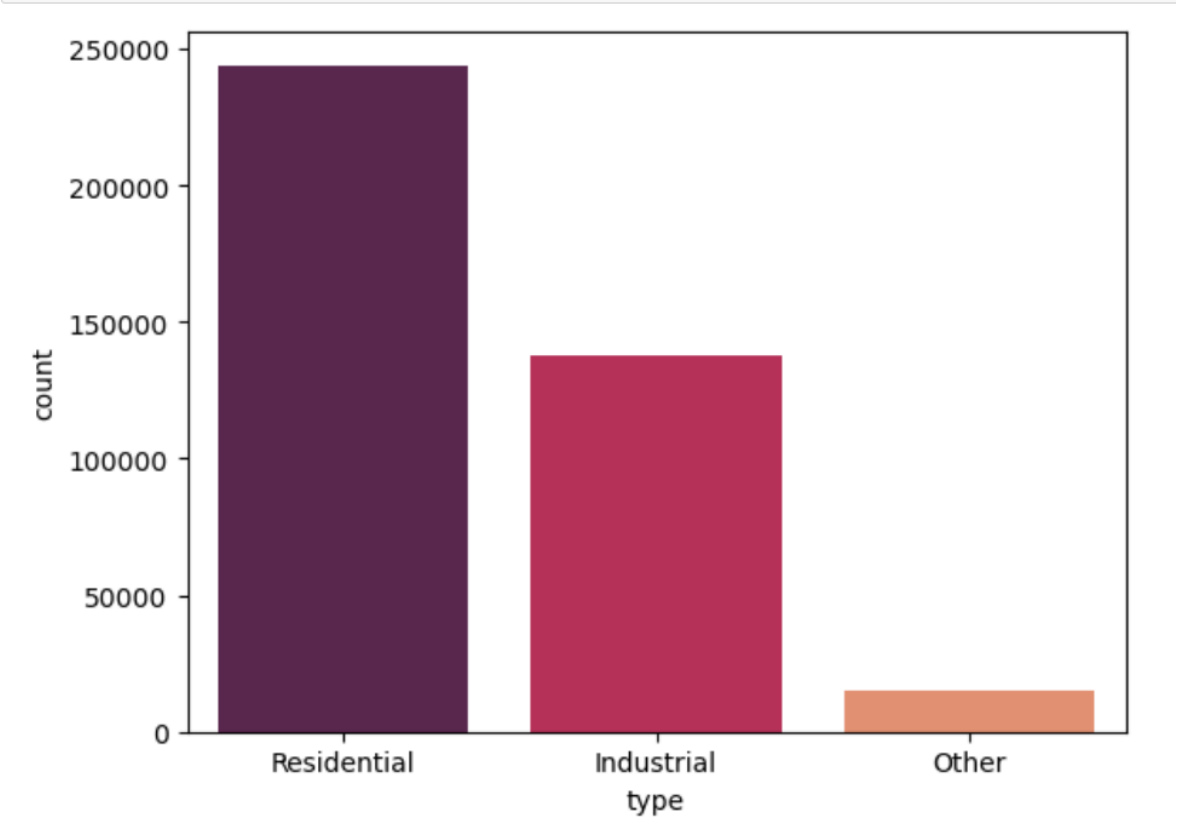
Drishti	Data Analysis, Synopsis
Abhishek Shukla	Data Analysis, Synopsis

Tables of Content

1. Introduction to Problem

Rapid urbanization and industrialization have led to an alarming increase in pollution levels, adversely affecting air, water, and soil quality. The lack of comprehensive data analysis tools hinders our ability to understand the dynamic patterns and sources of pollution. This project aims to address this gap by conducting a thorough analysis of pollution data, identifying key contributors, and developing effective strategies for pollution control and mitigation. The goal is to provide actionable insights to policymakers and communities, fostering informed decision-making for a sustainable and healthier environment.

2. E-R Model



3. Requirements

3.1 Technology Stack

Python: High-level programming language used for server-side scripting.

Jupyter Notebook: Jupyter Notebook is an open-source web application that allows you to create and share documents containing live code, equations,

visualizations, and narrative text, providing an interactive and collaborative environment for data science and analysis.

3.2 Hardware

Laptop/ Computer

3.3 Software

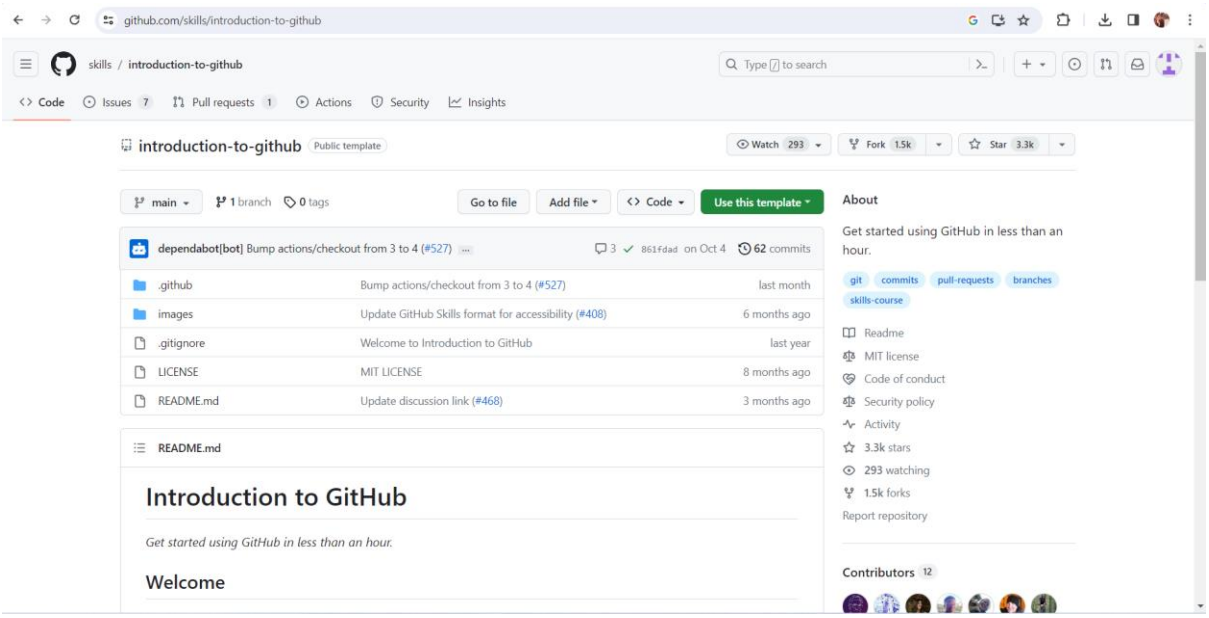
Operating System (OS)

Version Control System

Text Editors and Integrated Development Environments (IDEs)

3.4 Deployment Environment

Github



4. Overview

The data analysis project aims to investigate and derive meaningful insights from a specific dataset. It involves collecting, cleaning, and processing raw data to uncover patterns, trends, and correlations. Using statistical methods and visualization tools, the project seeks to provide a comprehensive understanding of the data, enabling informed decision-making. The analysis may involve exploring relationships between variables, identifying outliers, and creating predictive models. Throughout the project, a systematic approach is followed, including hypothesis testing and validation of results. The ultimate goal is to offer actionable recommendations or conclusions based on the data findings. The project typically employs programming languages such as Python or R, along with tools like Jupyter Notebooks, to facilitate a transparent and reproducible analytical workflow. Overall, the data analysis project

serves to extract valuable insights, enhance understanding, and support evidence-based decision-making in a given domain.

5. Project Module

1. Import the required libraries.
2. Load/ Read the Dataset
3. Prepare EDA
4. Do Visualizations
5. Effect of different gases on different states
6. Prepare Heatmap/ Confusion Matrix
7. Prepare Profile Report

6 Sample Screenshots

Data Analysis on Pollution

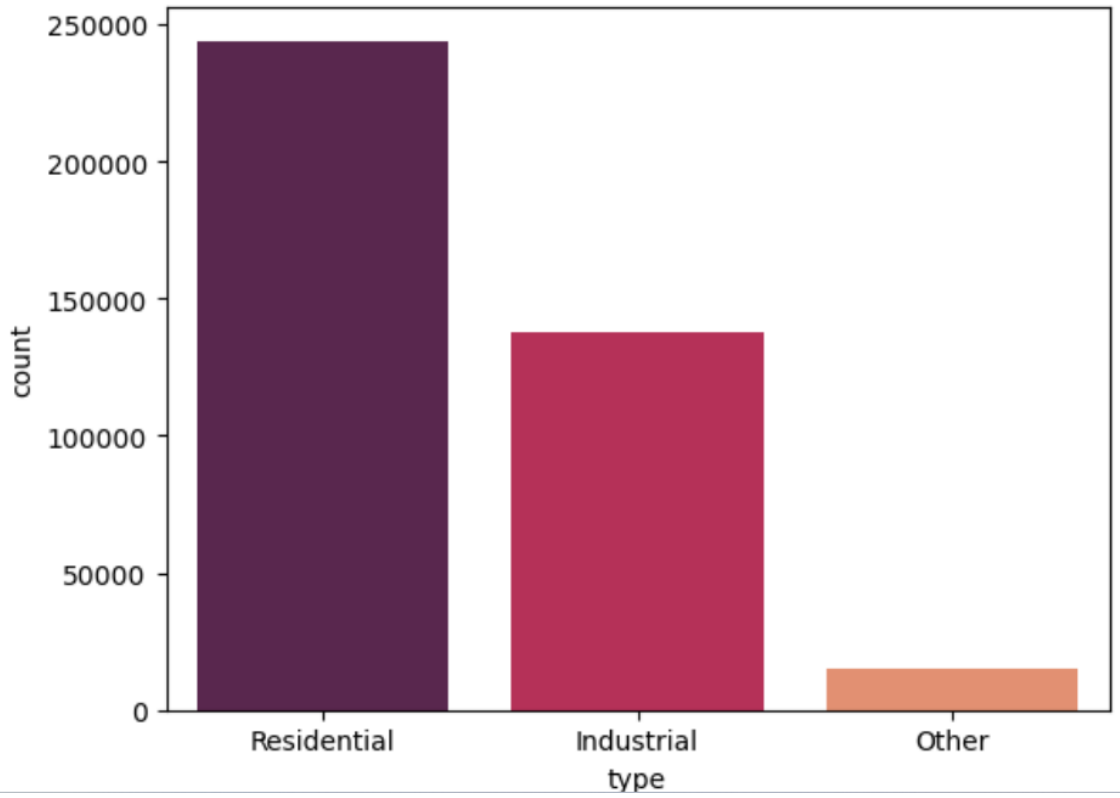
In [40]:
```
#import the required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import matplotlib
import seaborn as sns
%matplotlib inline
```

In [2]:
```
import warnings
warnings.filterwarnings('ignore')
```

In [3]:
```
#read the dataset
data = pd.read_csv('Pollution.csv')
```

Description of the Dataset is as follows:

1. stn_code : Station code. A code is given to each station that recorded the data.




```
In [38]: #import the another library
         from ydata_profiling import ProfileReport

In [39]: prof = ProfileReport(data)
         prof.to_file(output_file = 'output.html')
```

Summarize dataset: 100% 56/56 [00:35<00:00, 1.31it/s, Completed]

Generate report structure: 100% 1/1 [00:12<00:00, 12.37s/it]

Render HTML: 100% 1/1 [00:05<00:00, 5.19s/it]

Export report to file: 100% 1/1 [00:00<00:00, 21.72it/s]

7 Source Code

Pandas Profiling Report

Overview | Variables | Interactions | Correlations | Missing values | Sample | Duplicate rows

More details

location

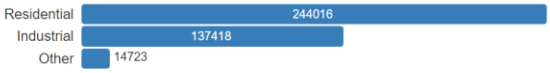
Text

Distinct	296
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	6.0 MiB

type

Categorical

Distinct	3
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	6.0 MiB



More details

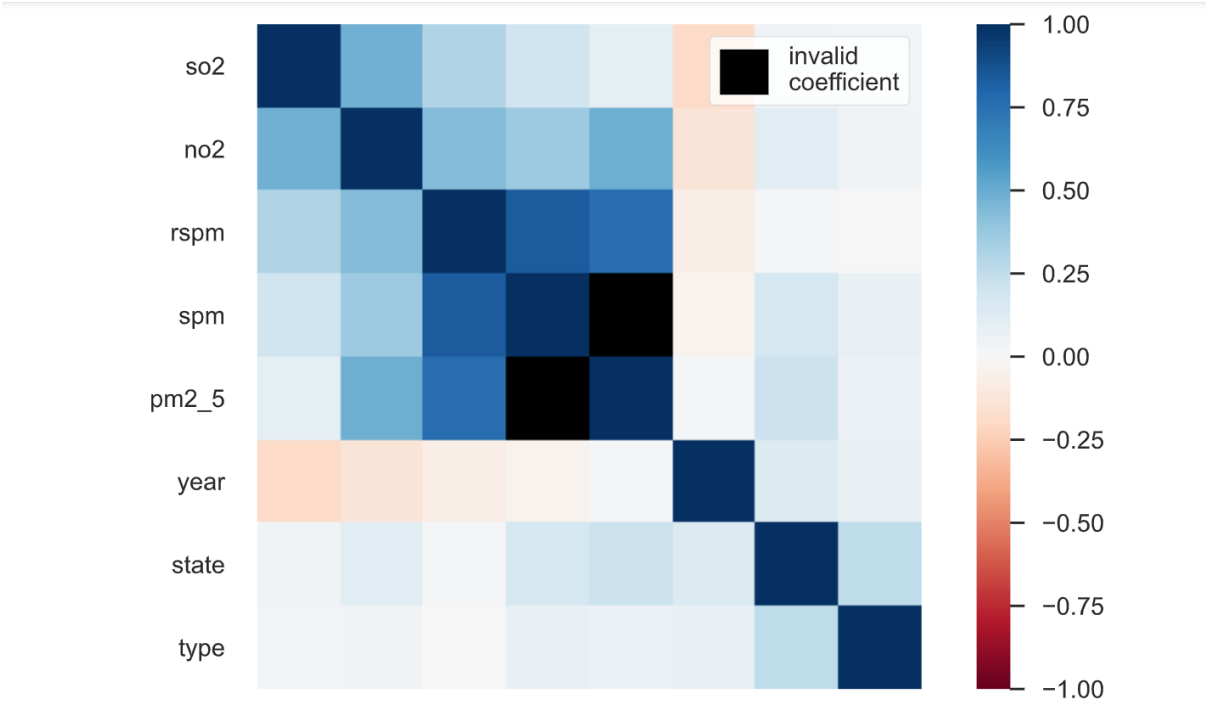
rspm

Real number (ℝ)

HIGH CORRELATION MISSING

Distinct	5823	**Minimum**	0
Distinct (%)	1.6%	**Maximum**	6307.0333
Missing	29643	**Zeros**	565
Missing (%)	7.5%	**Zeros (%)**	0.1%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	110.37268	**Memory size**	6.0 MiB

More details



8 Future Scope

With the increasing volume and variety of data generated, the future will likely see a greater emphasis on big data analytics, exploring large datasets to extract meaningful patterns and insights.

As the need for instant insights grows, real-time data analysis will become more prominent, especially in industries like finance, healthcare, and IoT (Internet of Things).

9 Conclusion

In conclusion, this data analysis project has successfully unveiled valuable insights, revealing patterns and trends within the dataset. The systematic exploration of relationships between variables has provided a deeper understanding of the underlying dynamics. The findings offer a foundation for informed decision-making, guiding future strategies and actions. The project's use of advanced analytical tools and methodologies showcases the evolving landscape of data science. Moving forward, continuous advancements in machine learning, artificial intelligence, and big data analytics will shape the future of data analysis. Ethical considerations must remain at the forefront to ensure responsible data usage and unbiased results. Collaboration between data scientists

and domain experts will further refine analyses for specific industries. The project highlights the importance of transparency and reproducibility in analytical workflows for fostering trust in results. As we embrace emerging technologies, the scope for data analysis remains expansive, promising innovative solutions to complex challenges across diverse domains.

10 References

https://www.kaggle.com/datasets

THANK YOU