IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Review Of Exploratory Data Analysis Techniques With Application To Air Quality Analysis

[1]Deepanjan Das, [2]Kanhaiya Kumar, [3]Abhay Kumar, [4]Hans Raj

[1]Student, [2]Student, [3]Student, [4]Student

[1]Computer Science and Engineering,

[1]Parul University, Vadodara, India

Abstract: Exploratory Data Analysis (EDA) is a fundamental step in the data analysis process that helps in understanding the structure and characteristics of data. It involves techniques such as statistical summaries, data visualization, correlation analysis, and anomaly detection. This paper presents a review of various EDA techniques and discusses their importance in extracting meaningful insights from datasets. Furthermore, the application of EDA in air quality analysis is explored, where it helps identify pollution trends, seasonal variations, and relationships between pollutants. The study highlights how EDA supports decision-making and improves the effectiveness of further analytical methods, including machine learning.

Keywords **-** Exploratory Data Analysis, Data Visualization, Air Quality, AQI, Time-Series Analysis.

1.INTRODUCTION

With the increasing availability of data, analysing and understanding datasets has become very important. Before using advanced methods like machine learning, it is necessary to explore and understand the data properly. Exploratory Data Analysis (EDA) helps in this process by providing simple techniques to examine and summarize data.

EDA allows researchers to identify patterns, detect errors, and understand relationships between variables. It serves as the first step in the data analysis process and helps in preparing data for further analysis.

One of the major areas where EDA is applied is air quality analysis. Due to rising pollution levels, analysing air quality data is important for environmental and health-related decisions. EDA techniques help in studying pollution patterns and understanding changes in air quality over time.

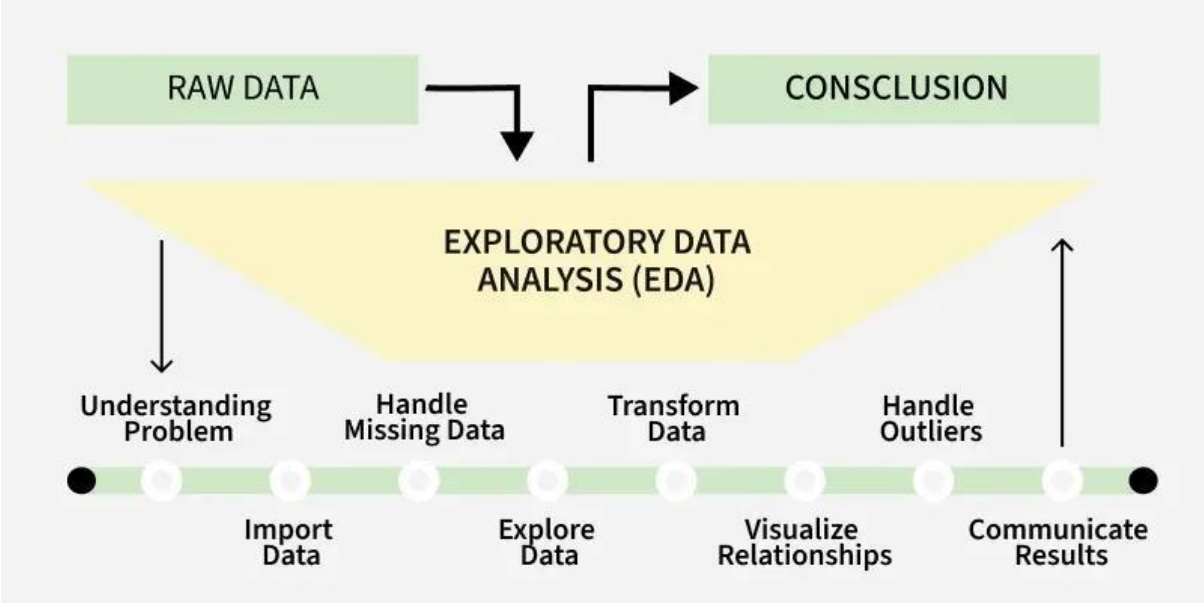

Figure 1: Steps involved in the Exploratory Data Analysis (EDA) process

2. LITERATURE REVIEW

Exploratory Data Analysis (EDA) has been widely recognized as a fundamental step in the data analysis process. Various researchers have worked on its development and application across multiple domains.

Chatfield (2011) described EDA as an essential initial step in statistical analysis. The study emphasized the use of graphical techniques and summary statistics to understand the structure of data before applying complex models. It highlights that visualization plays a key role in identifying patterns and trends.

Komorowski et al. (2016) explained that EDA is useful in examining data distribution, detecting outliers, and identifying anomalies. The authors pointed out that these techniques help improve data quality and ensure more accurate analysis. Their work also emphasized the importance of understanding data before performing predictive modeling.

Morgenthaler (2009) discussed the theoretical foundations of EDA and its importance in statistical practice. The study builds upon the work of John Tukey, who introduced EDA as a systematic approach to analysing data. It highlights that EDA is not only about visualization but also about developing intuition regarding the dataset.

Ramyasri and Raju (2025) applied EDA techniques along with data visualization and machine learning to analyse air quality data. Their research demonstrated that EDA can effectively identify pollution patterns and provide meaningful insights into environmental data. The study also showed that combining EDA with machine learning improves the overall analysis process.

Yu (2016) focused on the use of EDA in time-series air quality data. The study utilized visualization and temporal analysis techniques to understand variations in pollution levels over time. It highlighted how seasonal patterns and trends can be identified using time-series analysis.

Dubes and Jain (1980) explored clustering methodologies within the context of EDA. Their study explained how grouping similar data points can help simplify complex datasets and reveal hidden patterns. Clustering is particularly useful when dealing with large datasets where manual analysis is difficult.

In addition to these studies, several researchers have emphasized the importance of combining multiple EDA techniques to achieve better results. Visualization, statistical analysis, and clustering together provide a more comprehensive understanding of the dataset.

Overall, the literature suggests that EDA is a versatile and powerful tool that plays a critical role in both theoretical and practical data analysis. Its application in domains such as air quality analysis highlights its importance in solving real-world problems.

---

3. EXPLORATORY DATA ANALYSIS TECHNIQUES

3.1 STATISTICAL ANALYSIS

Statistical measures such as mean, median, and standard deviation help summarize data and understand variability.

3.2 DATA VISUALIZATION

Visualization techniques such as histograms, line graphs, and scatter plots help in identifying patterns and trends.

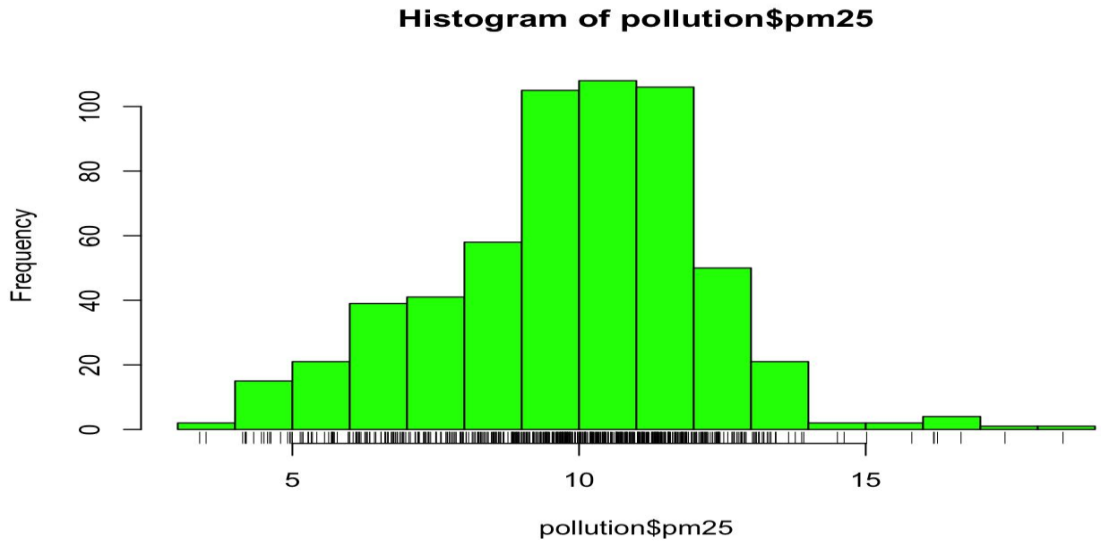

Figure 2: Histogram showing the distribution of air quality data values

3.3 OUTLIER DETECTION

Outliers are unusual values that may affect analysis and need to be handled carefully.

3.4 CORRELATION ANALYSIS

Correlation helps identify relationships between variables.

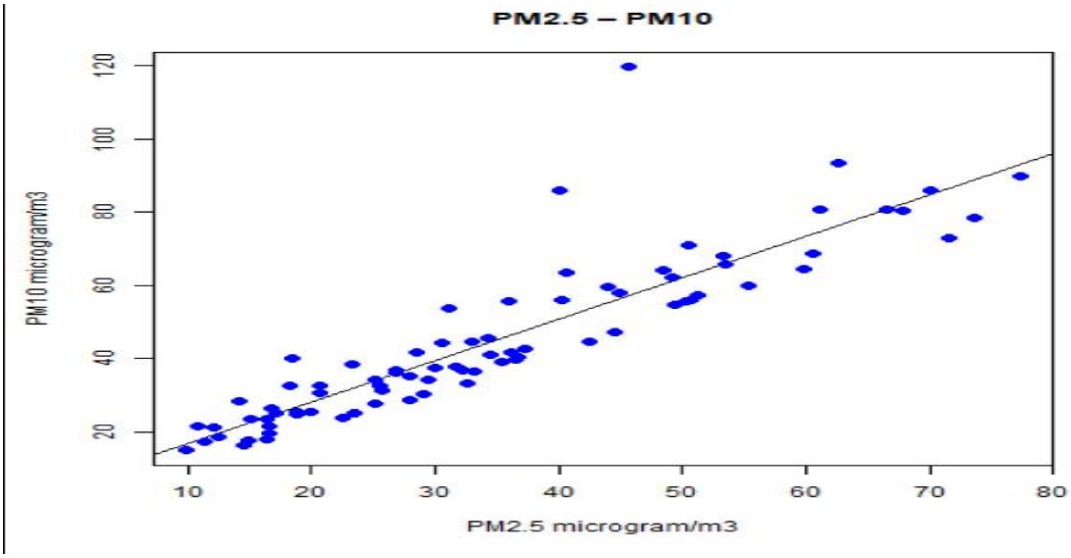

Figure 3: Scatter plot showing correlation between PM2.5 and PM10 pollutant levels

3.5 TIME-SERIES ANALYSIS

Used to analyse data over time and identify trends and seasonal patterns.

3.6 CLUSTERING

Clustering groups similar data points and simplifies complex datasets.

4. APPLICATION IN AIR QUALITY ANALYSIS

Air quality analysis involves studying pollutants such as AQI, PM2.5, PM10, CO, and NO_2. EDA helps identify pollution trends, seasonal variations, and relationships between pollutants.

Visualization techniques help in representing pollution data clearly. Time-series analysis helps in understanding long-term patterns. These insights are useful for environmental monitoring and decision-making.

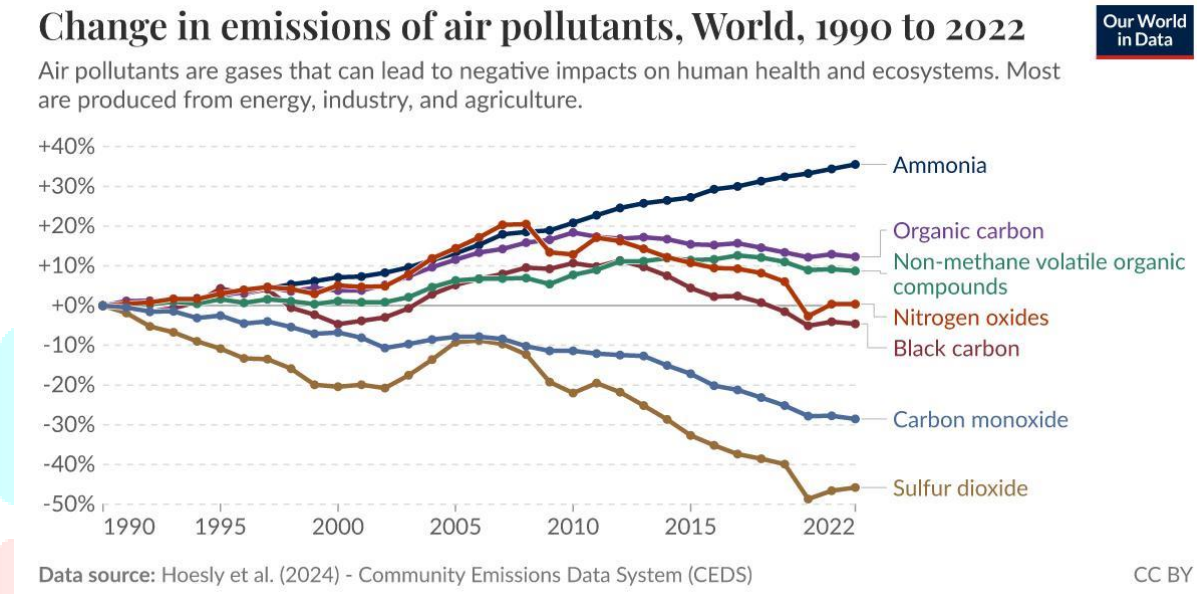

Figure 4: Time-series graph showing variation in air quality over time

5. LIMITATIONS

1. EDA does not provide predictive results
2. Depends on data quality
3. Large datasets require advanced tools
4. Interpretation may vary

6. FUTURE SCOPE

EDA can be improved by integrating machine learning and automation techniques. In air quality analysis, real-time monitoring and prediction systems can be developed for better environmental management.

7. CONCLUSION

EDA is an essential step in data analysis that helps in understanding datasets and extracting meaningful insights. This paper reviewed EDA techniques and their application in air quality analysis, showing their importance in real-world problems. EDA provides a strong foundation for further advanced techniques such as machine learning and predictive analytics.

8. REFERENCES

[1] C. Chatfield, *Exploratory Data Analysis*, 2011.

[2] M. Komorowski et al., *Exploratory Data Analysis*, 2016.

[3] S. Morgenthaler, *Exploratory Data Analysis*, 2009.

[4] M. N. Ramyasri and A. N. Raju, *Analysis of Air Quality using EDA and Machine Learning*, 2025.

[5] C. Yu, *Research of Time Series Air Quality Data Based on EDA*, 2016.

[6] R. Dubes and A. K. Jain, *Clustering Methodologies in Exploratory Data Analysis*, 1980.