Predictive Analysis of Toxic Releases

- **Title:** Predictive Analysis of Toxic Releases in the United States (2024-2028)
- **Prepared by:** Madhumitha Somasundaram, Sathish Kumar Prabaharan
- **Date:** 11-22-2024
- **Affiliation:** University of Colorado,Boulder

TABLE OF CONTENTS :

ABSTRACT

This report presents a predictive analysis of toxic waste releases in the United States, utilizing data from the EPA's Toxics Release Inventory (TRI) program, spanning from 2014 to 2023. The study aims to identify trends and patterns in toxic waste emissions to assist in regulatory compliance and improve environmental protection efforts. The dataset includes various factors such as facility details, geographical information, industry sectors, and the types and quantities of chemicals released. By cleaning, transforming, and analyzing this data, the goal is to uncover key patterns contributing to toxic waste releases and inform strategies to reduce environmental hazards. This report will also evaluate various predictive models to forecast toxic waste releases and explore their implications for environmental policy and management.

INTRODUCTION

The release of toxic waste into the environment has significant implications for public health and ecosystem stability. The " Predictive Analysis Of Toxic Releases" project aims to develop a system that uses historical TRI data to predict future toxic release patterns across industries and regions. By analyzing trends in toxic emissions, the system will identify potential future risks, helping government agencies, environmental organizations, and industries make informed decisions to reduce environmental harm, comply with regulations, and promote sustainability. Key factors such as facility

data, industry sectors, types of chemicals, and release quantities will be analyzed to uncover patterns and build a predictive model. The outcome of this project will provide actionable insights for proactive environmental management, assist in regulatory compliance, and contribute to public health protection. Ultimately, the goal is to transform TRI data into a comprehensive tool for minimizing the impact of toxic releases and fostering a healthier, more sustainable environment.

DATASET OVERVIEW

The dataset used in this project consists of historical toxic waste release records from various facilities in the United States, sourced from the EPA's Toxics Release Inventory (TRI). It includes a range of attributes relevant to predicting and analyzing toxic waste emissions, such as:

YEAR: Details of the year from 2014 to 2023.

FACILITY NAME: The name of the facilities in the U.S. that report toxic waste releases.

CITY: The cities in the U.S. where the facilities are located.

ST: The states in the U.S. where the facilities are located.

INDUSTRY SECTOR: The type of industry that is responsible for releasing chemicals.

CHEMICAL: The types of chemicals released from the facility.

CARCINOGEN: Indicates whether the released chemicals are carcinogenic.

WATER: Chemicals released via water bodies.

OTHER DISPOSAL: Other forms of chemical disposal used by the facility.

ON-SITE RELEASE TOTAL: The total amount of chemicals released on-site.

OFF-SITE RELEASE TOTAL: The total amount of chemicals released off-site.

TOTAL TRANSFER: The total off-site activities, including releases and treatments conducted off-site.

TOTAL RELEASES: The total amount of chemicals released both on-site and off-site.

AIR RELEASES: Chemicals released via the air.

LAND RELEASES: Chemicals released to land.

TOTAL ON-SITE: The total amount of on-site activities, including releases and treatments conducted on-site.

ON-SITE OTHER TREATMENT: Includes other on-site treatment activities such as energy recovery, recycling, etc.

OFF-SITE OTHER TREATMENT: Includes other off-site treatment activities such as energy recovery, recycling, etc.

TOTAL OTHER TREATMENT: The total amount of other treatment activities conducted both on-site and off-site.

This dataset spans several years (2014 to 2023) and provides comprehensive details on toxic waste emissions from a large number of facilities across various U.S. cities and states. It includes information on both on-site and off-site releases, as well as treatment activities, making it suitable for both predictive modeling and regulatory compliance analysis. By analyzing the data, the project seeks to uncover trends and patterns in toxic waste releases, providing insights into the impact of these releases on public health and the environment.

DATA COLLECTION :

The dataset was collected from the U.S. Environmental Protection Agency (EPA), which provides data from the Toxics Release Inventory (TRI). The TRI is a resource for learning about toxic chemical releases and pollution prevention activities reported by industrial and federal facilities. This dataset spans several years and includes comprehensive records on the release of toxic chemicals into the environment, providing insights into pollution patterns across various industries and regions. The data collected from TRI will be used to analyze trends, evaluate environmental risks, and develop predictive models for future toxic waste releases.

DATA EXPLORATION :

We performed several data preparation and cleaning steps to make the dataset accurate, consistent, and ready for the analysis of toxic releases.These steps included addressing missing values, merging relevant datasets, transforming variables for better analysis, and conducting thorough data quality checks. The goal was to improve the dataset's reliability and make it suitable for identifying patterns and trends in toxic chemical releases, enabling more effective predictive modelling and environmental impact assessments.

DATA INTEGRATION :

The dataset comprises toxic release records from 2014 to 2023, initially stored in separate CSV files. To streamline the data collection and analysis process, we automated the code to download the yearly CSV files directly from the EPA website. Once the files were downloaded, we loaded each year's data into DataFrames. Afterward, we combined these DataFrames into a single, unified DataFrame. This integration involved merging the yearly DataFrames for 2014 through 2023, and then exporting the combined DataFrame into a single consolidated CSV file. This approach facilitated more efficient analysis, as it allowed us to work with a cohesive dataset that could be easily filtered, grouped, and analysed across the entire time span, providing valuable insights into toxic release trends and patterns over the years.

DATA CLEANING :

Data cleaning is an essential step to make sure the dataset is accurate, reliable, and ready for analysis. It involves fixing errors, resolving inconsistencies, and filling in missing information. Here are the main steps we followed:

- **Removing Duplicates:** Duplicate records, where facilities had reported the same data multiple times, were identified and removed to ensure each entry was unique and accurate.
- **Handling Missing Data：**
    1. **For Categorical Columns**: Missing values are filled with "Unknown" to maintain categorical integrity.
    2. **For Numerical Columns**: Missing values are replaced with the mean of the respective column to avoid introducing bias.
    3. **For Carcinogen Column :** The missing values in the 'Carcinogen' column were filled by reviewing the classification of similar chemicals. Based on this assessment, the chemicals were categorised as either carcinogenic or non-carcinogenic, and the missing values were updated accordingly.
- **Renaming Columns：** Columns with long or complex names are renamed to more intuitive and concise labels (e.g., "8.1A - ON-SITE CONTAINED" becomes "ON-SITE CONTAINED"). This makes the dataset easier to read and work with.
- **Standardising Units :** The TRI data uses different measurement units (e.g., pounds and grams). To standardise:
    1. Any data reported in grams is converted to pounds for consistency.
    2. This ensures that numerical analysis and comparisons are accurate.

DATA REDUCTION :

The dataset originally contained 122 columns per year. To streamline the analysis and focus on meaningful attributes, we reduced the dataset to around 19 columns by removing non-essential data. Additionally, columns with redundant or related information were merged to enhance clarity and simplify the dataset. For example:

- **'AIR RELEASES'** was created by combining the **'FUGITIVE AIR'** and **'STACK AIR'** columns. This allowed for a more comprehensive representation of airborne chemical releases.
- **'LAND RELEASES'** was generated by combining the **'UNDERGROUND'**, **'LANDFILLS'**, **'LAND TREATMENT'**, and **'SURFACE IMPNDMNT'** columns. This provides a consolidated view of all land-related chemical releases.
- **'TOTAL ON-SITE'** was created by combining several on-site activities: **'ON-SITE RELEASE TOTAL'**, **'TREATMENT ON SITE'**, **'ENERGY RECOVER ON'**, and **'RECYCLING ON SITE'**. This combined metric represents the overall activities taking place at the facility site.
- **'ON-SITE OTHER TREATMENT'** was formed by combining **'ENERGY RECOVER ON'**, **'RECYCLING ON SITE'**, and **'TREATMENT ON SITE'**, which reflect the on-site treatment activities such as energy recovery and recycling.
- **'OFF-SITE OTHER TREATMENT'** was formed by combining the treatment activities occurring off-site: **'ENERGY RECOVER OF'**, **'RECYCLING OFF SIT'**, and **'TREATMENT OFF SITE'**.
- **'TOTAL OTHER TREATMENT'** was created by summing the relevant columns associated with both on-site and off-site treatment activities. This provides a combined total of all treatment actions for the dataset.

Removed columns that are not required for further analysis. These columns are deemed unnecessary because they might contain redundant, irrelevant, or overly specific data.

DATA TRANSFORMATION :

DIFFERENCING :

To ensure stationarity in our time series data, we applied differencing specifically to the TOTAL RELEASES column. This column represents the primary metric under analysis, and making it stationary is crucial for accurate time series modelling. Differencing involves computing the difference between consecutive observations, effectively removing trends or seasonality from the

data.By differencing the TOTAL RELEASES column, we ensured that the series meets the stationarity requirement. This was validated using the **Augmented Dickey-Fuller (ADF) test**, confirming that the processed data is suitable for further analysis and modeling.

DATA ENCODING :

Categorical columns, such as 'CARCINOGEN', were encoded using label encoding to prepare the dataset for further analysis.

STATISTICAL ANALYSIS :

STATISTIC	TOTAL RELEASES	TOTAL OTHER TREATMENT
Count	**674,021**	**674,021**
Mean	**51,529.30**	**398,605.50**
Standard Deviation (Std)	**2,285,406**	**16,647,460**
Min	**0**	**0**
25th Percentile (25%)	**9**	**0**
Median (50%)	**327.71**	**1005**
75th Percentile	**4,896**	**31586**
Max	**678,057,400**	**3,754,222,000**

- **Min Values:** All columns have a minimum value of 0, indicating that some records report no data for these categories.

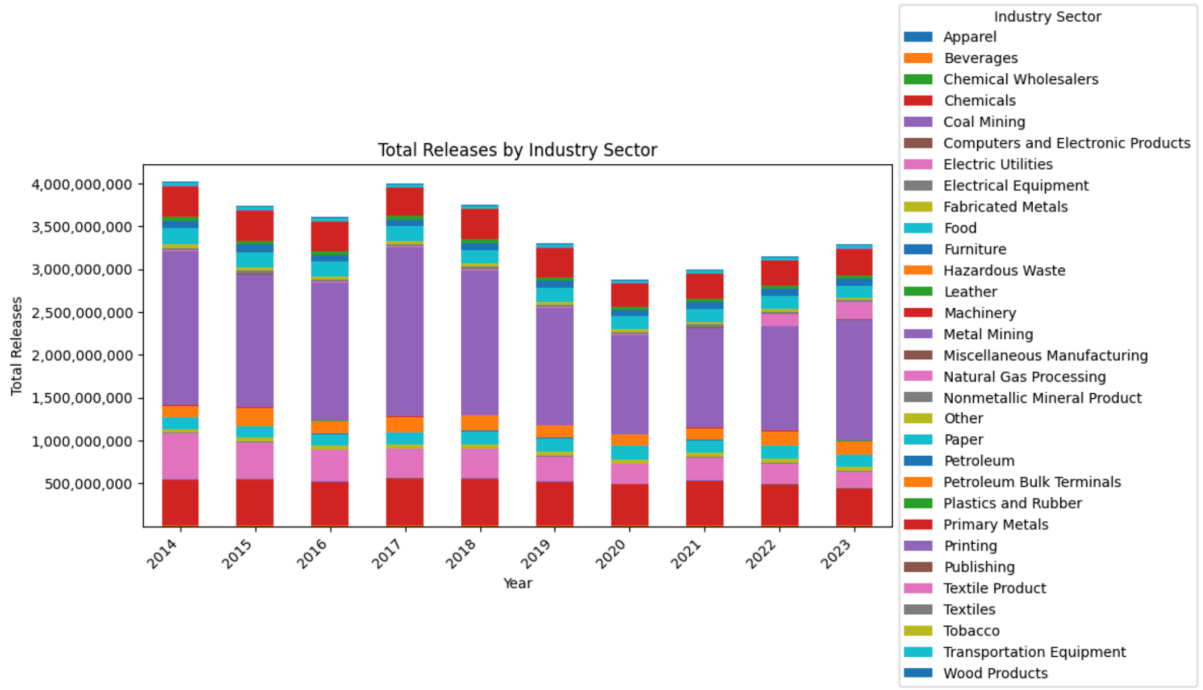
- **25th Percentile (0 values):** A large portion of the dataset has 9 values for these columns, indicating that for many records, few activity in these categories was reported.
- **Max Values:** The maximum values for these categories are very large, suggesting that some facilities report extremely high amounts for total treatment, transfer, releases, or on-site activities.
- **Standard Deviation:** The high standard deviation values indicate considerable variability in the data. Some facilities report very high values compared to others, contributing to the wide spread in the data.
- **Median (50th Percentile):** The median gives a better sense of the central tendency for these variables, with values generally lower than the mean, reflecting the skewed distribution due to a few very high values.

DATA VISUALIZATION :

Various visualisations were generated to explore trends in toxic releases over time, identify key sources of emissions, and examine distributions:

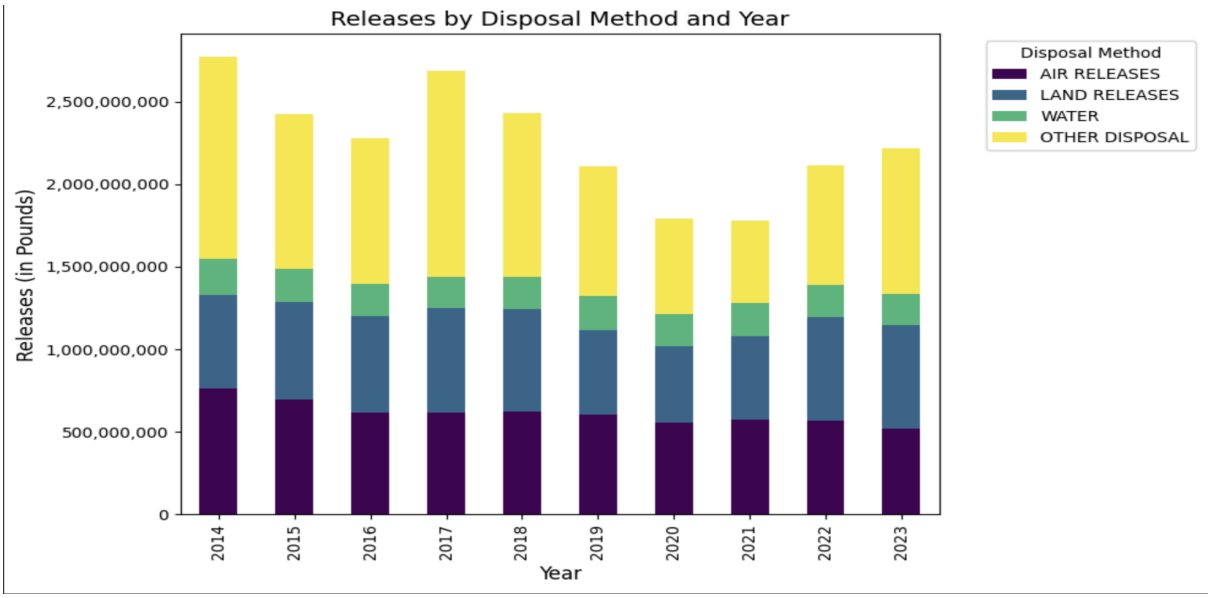
- **Time-Series Analysis**: Charts displaying annual release trends to reveal significant shifts or patterns in emissions.
- **High-emission Areas**: Illustrating toxic releases across regions to highlight high-emission zones.
- **Sectoral Emissions**: Bar graphs breaking down emissions by industry to identify major contributors to toxic releases.

1. TOTAL RELEASES BY INDUSTRY SECTOR :



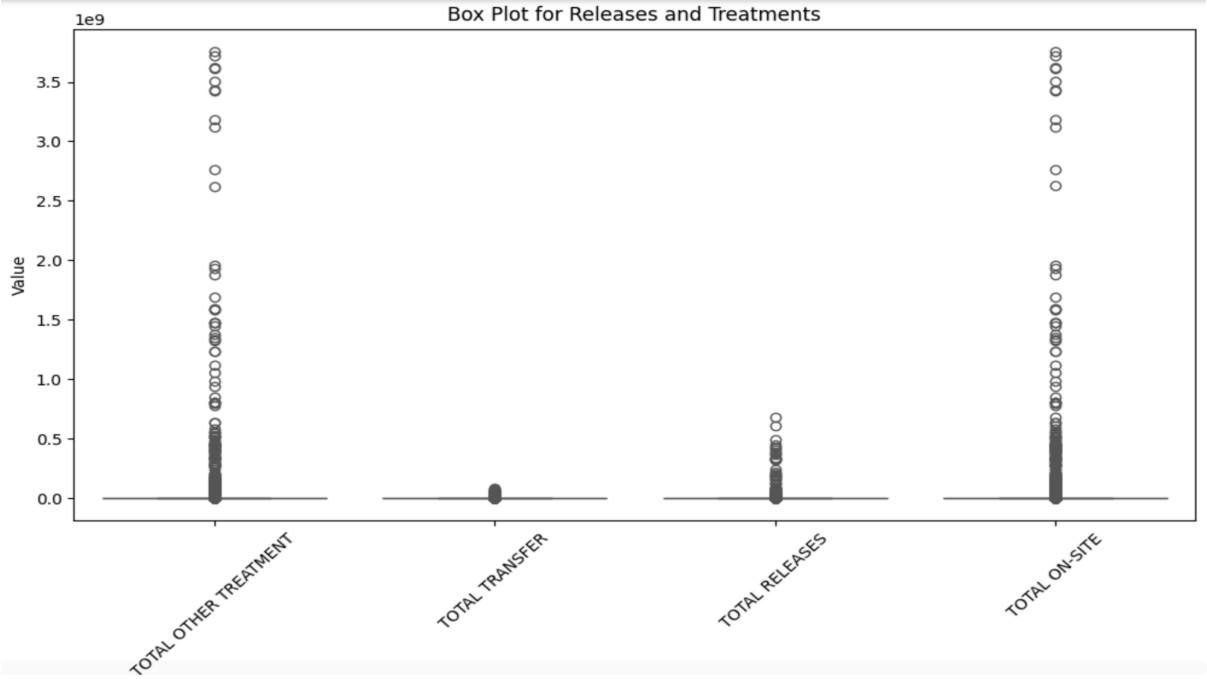
The total releases vary from year to year but are consistently high, ranging from 2.5 billion to 4 billion pounds.Different industry sectors contribute varying proportions to the total releases, with certain sectors like **Chemicals**, **Coal Mining**, and **Petroleum** showing significant shares across multiple years.The overall trend appears stable, with no dramatic increase or decrease across these years, although minor fluctuations can be observed.

2. RELEASES BY DISPOSAL METHOD AND YEAR :



The disposal method contributing the most to releases is Other Disposal, which consistently takes up the largest portion of the stacked bars.Air Releases have a steady but somewhat lower contribution in comparison to Other Disposal and Land Releases, though it remains substantial.Land Releases and Water Disposal also account for a significant portion of the total releases, with Water Disposal slightly increasing in recent years.Trends: From 2014 to 2023, the relative proportions of each disposal method appear stable, with Other Disposal continuing to dominate. There is no drastic shift in disposal methods over the years.

3.BOX PLOT FOR RELEASES AND TREATMENT:

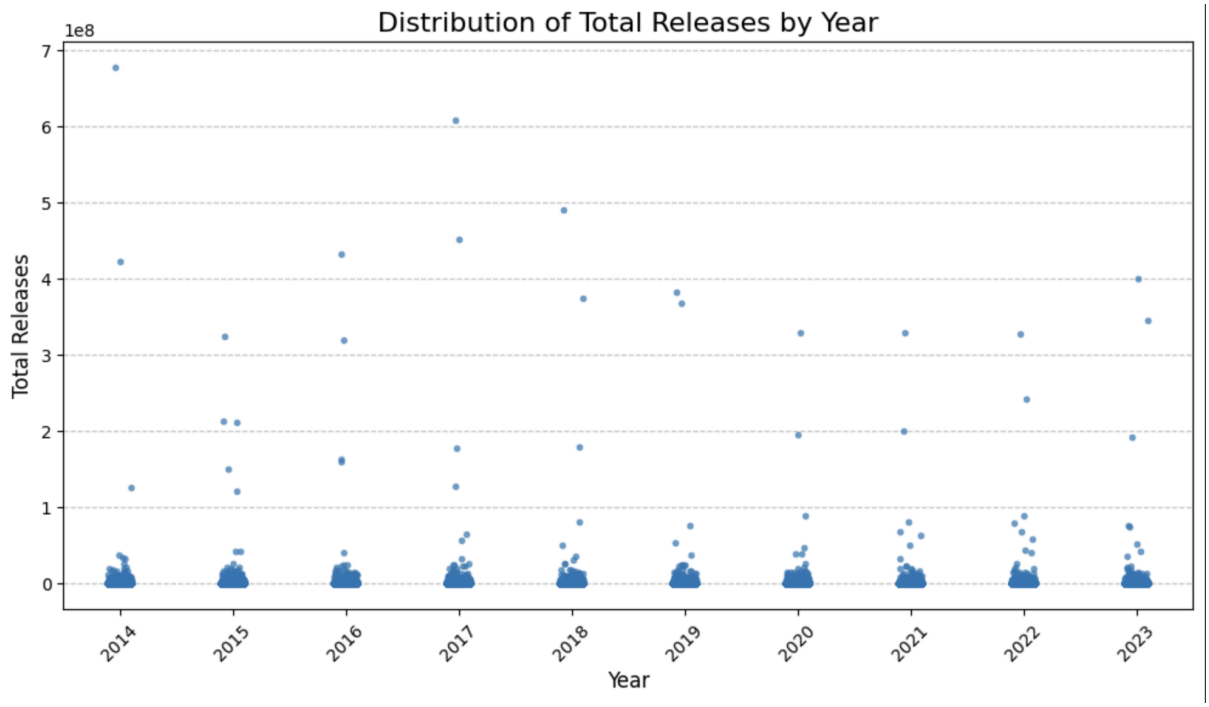


Most values in each category are concentrated near the lower range, with a few extreme outliers significantly inflating the scale.The extreme values suggest that a small number of incidents or sites contribute disproportionately to overall totals in some categories.The central tendency (medians) in all categories is close to zero, emphasizing that most values are small compared to the few outliers.

Category Differences:

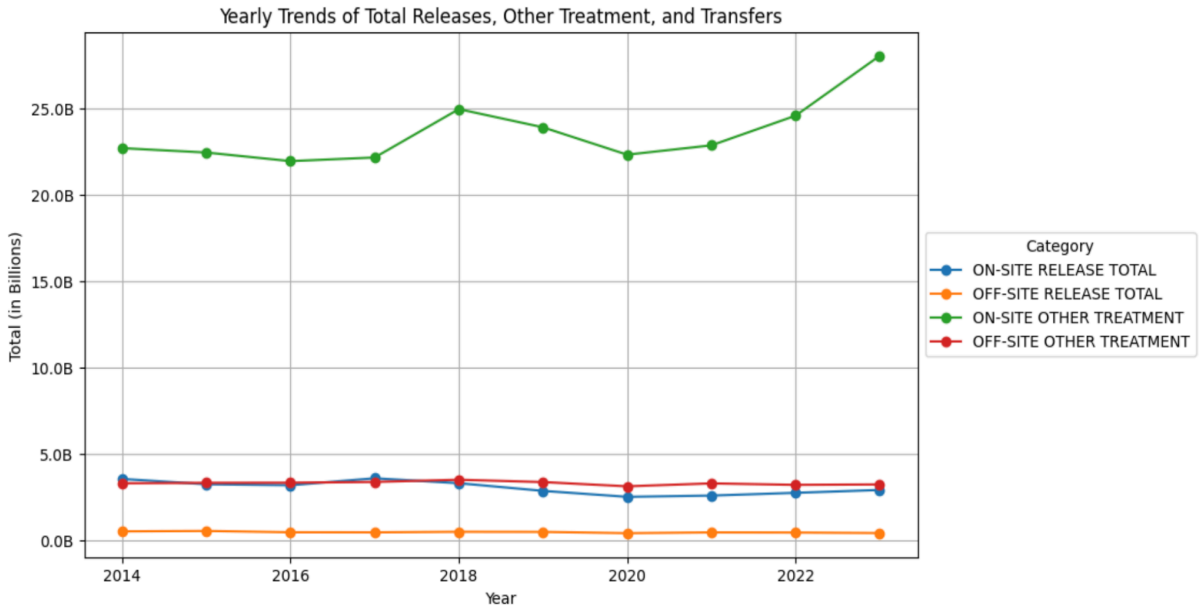
- "TOTAL ON-SITE" and "TOTAL OTHER TREATMENT" show the most significant outliers, indicating wide variability or occasional high-impact events.
- "TOTAL TRANSFER" has fewer outliers and appears more consistent compared to other categories.

4. DISTRIBUTION OF TOTAL RELEASES BY YEAR :



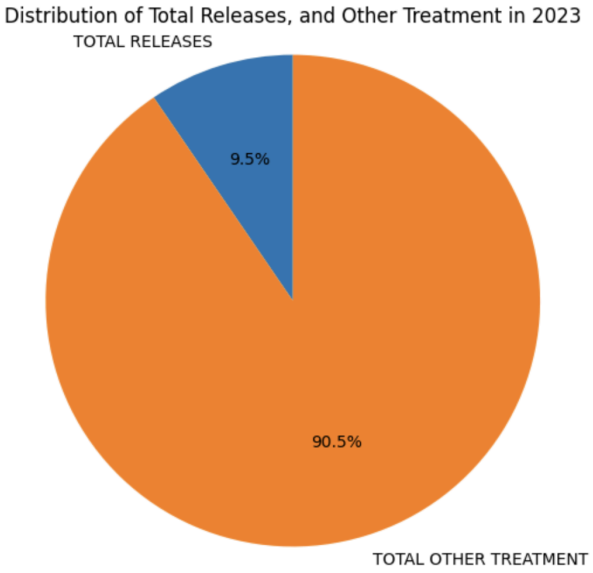
Certain years (e.g., 2015, 2017, 2020) have unusually high total release values, suggesting specific high-impact events or activities during those times.The majority of release values each year are near zero, showing that most events involve relatively minor releases.There is no consistent increase or decrease in total releases over the years, indicating that high-impact releases are sporadic rather than part of a sustained trend.

5. YEARLY TRENDS:



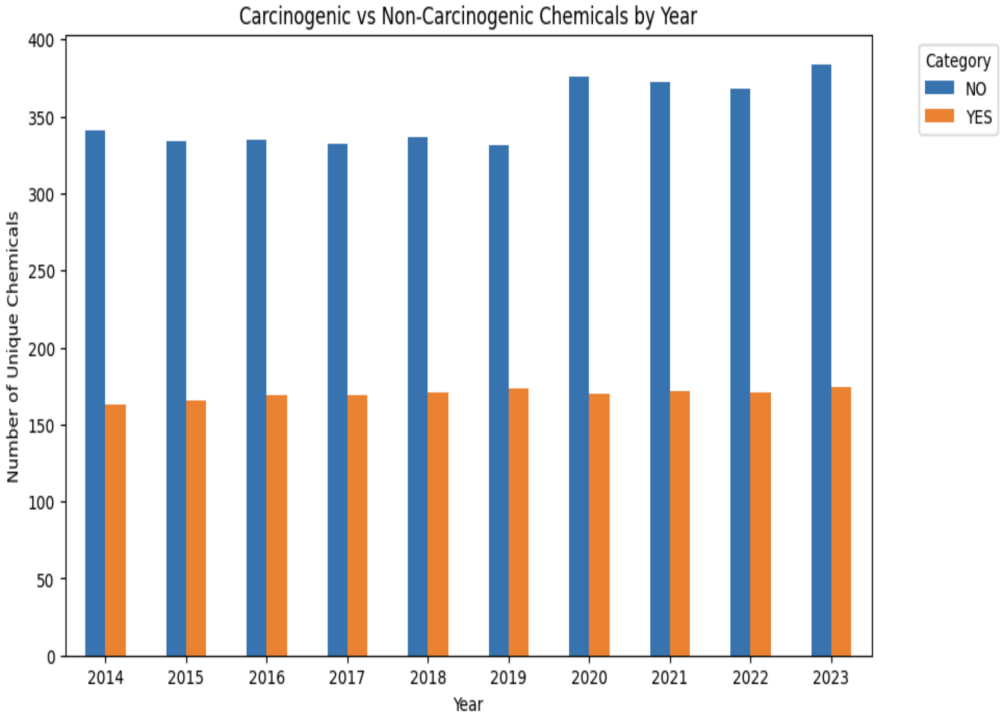
The on-site other treatment (green line) is significantly higher compared to other categories and has shown a steady increase over the years, particularly from 2020 to 2022.The on-site release total (blue line) and off-site other treatment (red line) remain relatively stable and much lower than the on-site other treatment.Off-site release total (orange line) is the lowest among all categories and does not exhibit any significant change over time.

6.DISTRIBUTION OF TOTAL RELEASES AND OTHER TREATMENT IN 2023 :



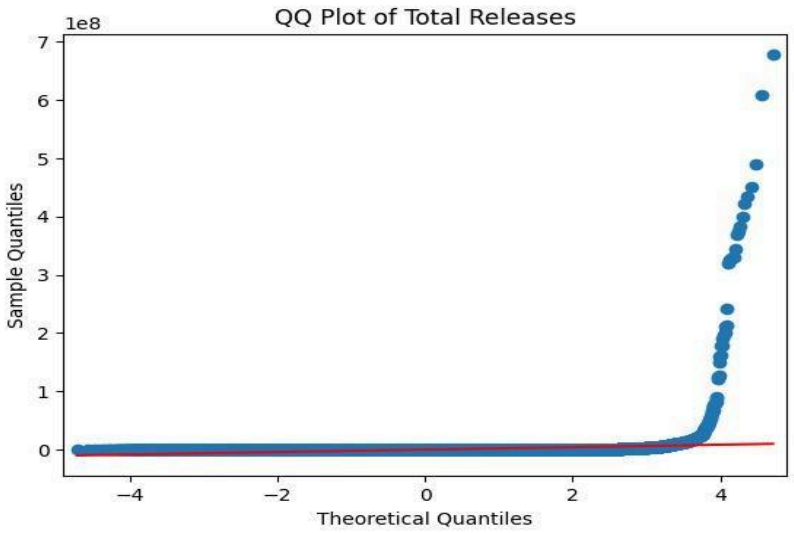
This chart shows the proportions of Total Releases versus Other Treatment in 2023.Total Releases indicating that only a small portion of waste or pollutants is being directly released.Other Treatment, which implies the majority of waste or pollutants undergo treatment instead of being directly released.A high proportion of waste being treated reflects a strong commitment to environmental sustainability and pollution management.It indicates effective waste treatment systems or policies in place, leading to reduced direct environmental releases.

7. CARCINOGENIC VS NON-CARCINOGENIC CHEMICALS OVER THE YEAR :



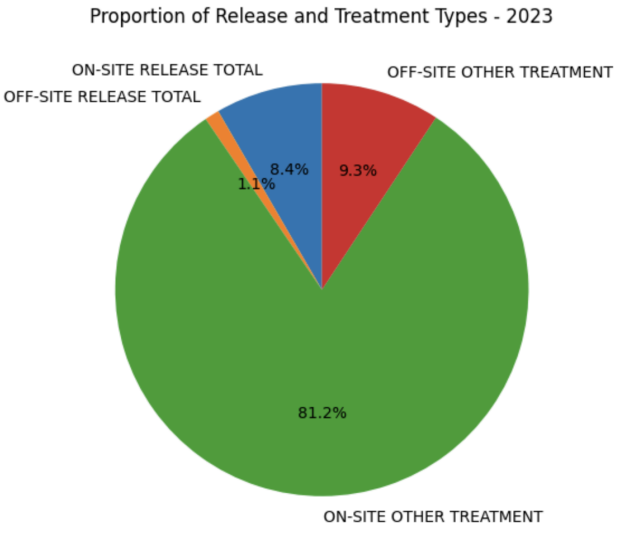
This bar chart compares the number of carcinogenic chemicals ("YES") and non-carcinogenic chemicals ("NO") across different years from 2014 to 2023.The count of non-carcinogenic chemicals (blue) is consistently higher than that of carcinogenic chemicals (orange) across all years.There appears to be little variation in the number of chemicals in each category from year to year.The count remains stable, ranging close to 350 chemicals each year. There is no apparent trend of increase or decrease over the years.On average, non-carcinogenic chemicals are more than twice as prevalent as carcinogenic ones.

8. Q-Q PLOT OF TOTAL RELEASES:



The QQ plot reveals that total chemical releases deviate from a normal distribution.Most points deviate from the red line, especially on the right side, indicating a few very high values (outliers) are affecting the data.

9.DISTRIBUTION OF RELEASES AND OTHER TREATMENT IN 2023 :

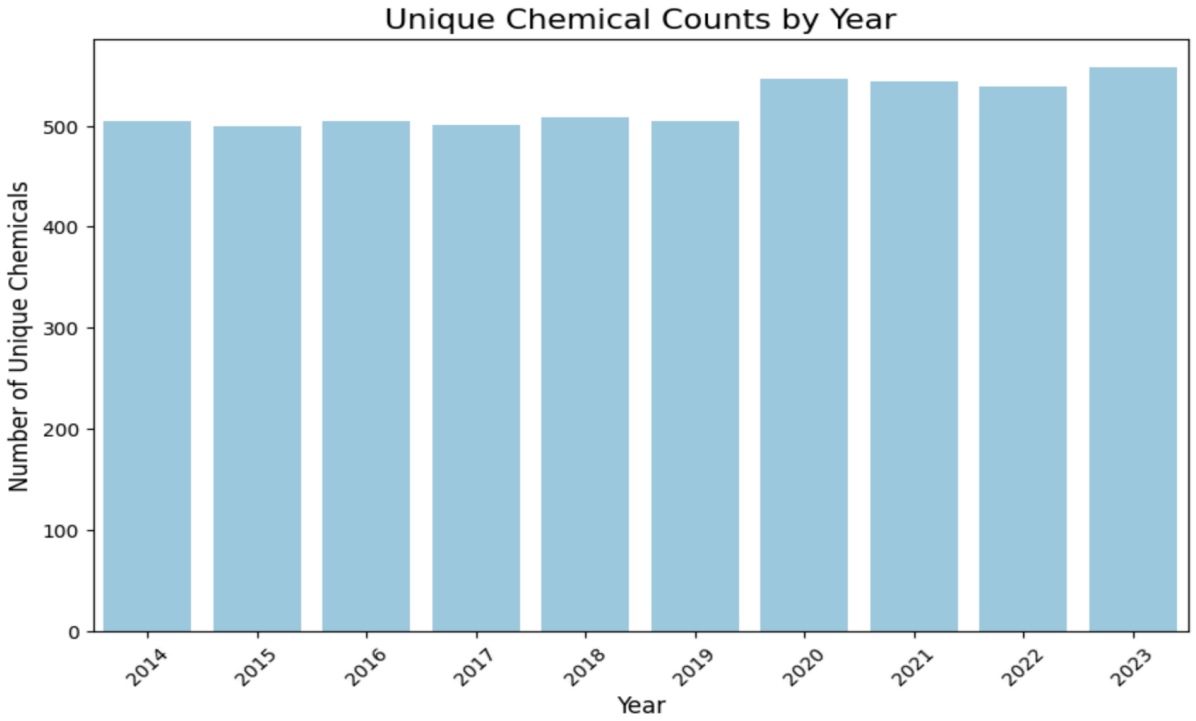


The chart segments the data into four distinct categories for how chemicals are managed:

- On-Site Other Treatment (81.2%): Majority of chemicals are treated on-site.
- Off-Site Other Treatment (9.3%): A smaller proportion undergoes off-site treatment.
- On-Site Release Total (8.4%): Represents chemicals released directly on-site.
- Off-Site Release Total (1.1%): The smallest share of chemicals are released off-site.

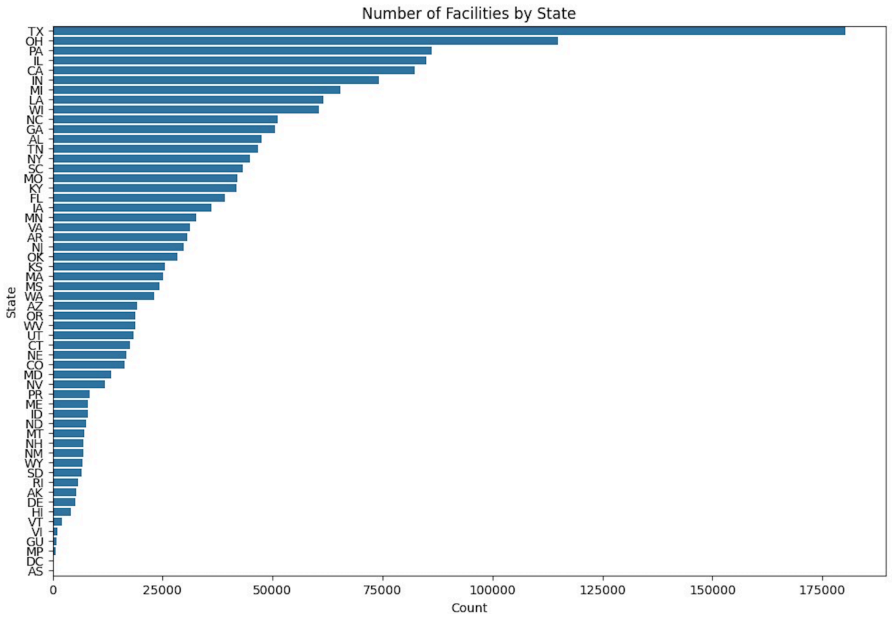
The majority of chemicals are managed via treatment processes, likely emphasizing containment or compliance with environmental regulations.Combined, chemical releases (both on-site and off-site) account for only 9.5% of the total.This suggests a strong focus on treatment over direct release into the environment.Indicates a preference for handling hazardous materials locally rather than transporting them, which could reduce the risk of contamination or accidents during transport.

10. RELEASE OF CHEMICALS OVER THE YEAR :



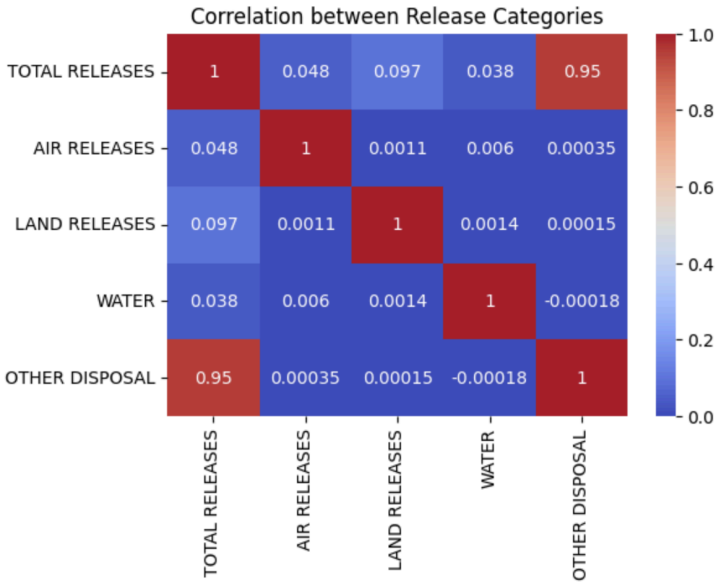
The number of unique chemicals reported each year is relatively consistent, hovering slightly above 500 chemicals annually.There is no significant increase or decrease over the years, suggesting a steady trend in the number of chemicals identified or tracked.Starting from 2020, there is a small but noticeable uptick in the total count of unique chemicals, culminating in the highest values in 2022 and 2023.This could indicate an expansion in chemical reporting or discovery during recent years.The chart implies that chemical monitoring efforts have been consistent over time, with no major shifts in the total number of chemicals being reported annually.

11.NUMBER OF FACILITIES ACROSS STATES IN US :



This chart displays the total number of facilities across different states. Texas has the highest number of facilities, while the average across other states is around 10,000 facilities per state.

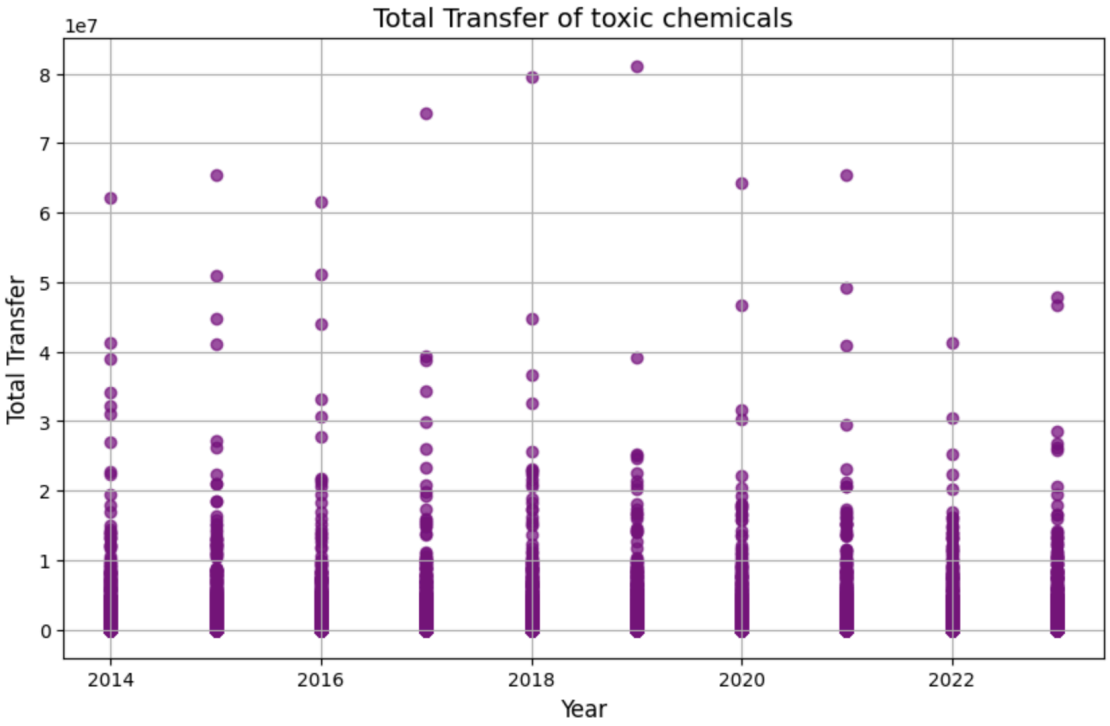
12. CORRELATION BETWEEN RELEASE CATEGORIES :



The heat map displays the pairwise correlation between five variables: Total Releases, Air Releases, Land Releases, Water Releases, and Other Disposal.The strength of correlation is colour-coded, with red representing strong positive correlations and blue representing weak or no correlations.The is a strongest correlation between Total Releases and Other Disposal (0.95).Indicates that "Other

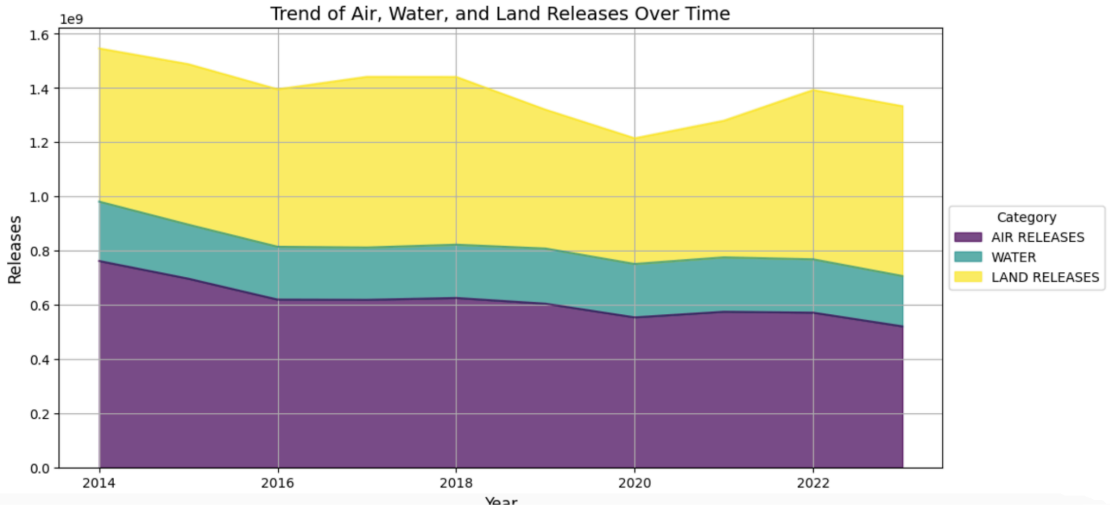
Disposal" is the major driver of overall releases.Minimal correlation between the other categories (e.g., air, land, water releases) suggests that their trends are largely independent of each other.

13.TOTAL TRANSFER OF TOXIC CHEMICALS OVER THE YEAR :



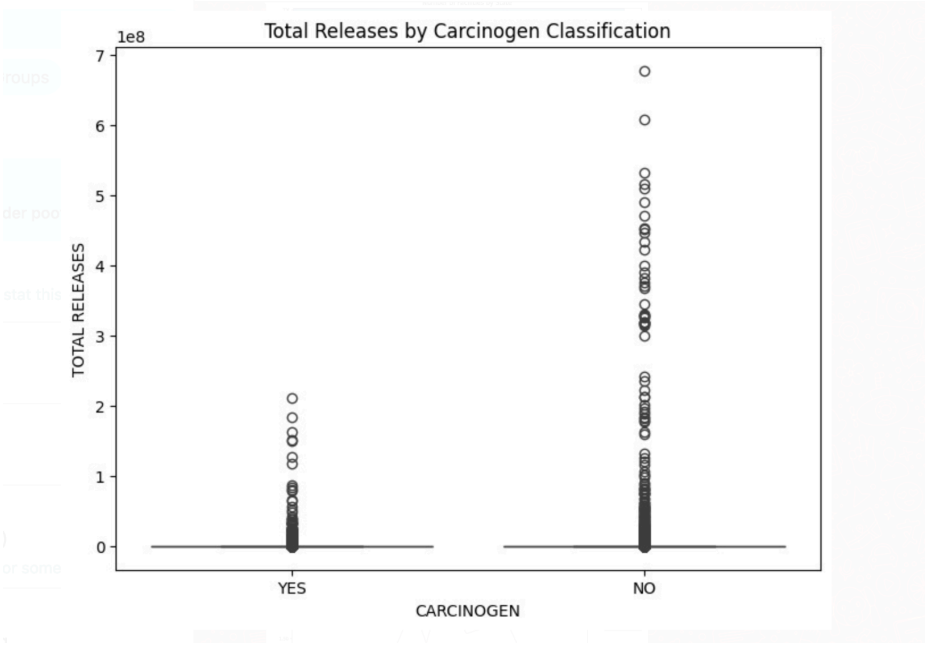
There is a wide distribution of transfer values across all years.The scatter plot highlights the presence of several outliers, with a few transfers reaching extremely high values .Most data points cluster around lower transfer values, indicating that the majority of transfers are smaller, with only a few instances of large-scale transfers.

14.TRENDS OF AIR, WATER AND LAND RELEASES OVER TIME :



Air Releases have a significant downward trend from 2014 to 2023 and it reflects success in reducing airborne pollutants.Water Releases are relatively stable, showing slight fluctuations without any dramatic changes and indicates either consistent efforts or stagnant focus on water pollution control.Land Releases are the largest contributor throughout the years.It has initially declined between 2014 and 2018 but increased again from 2019 onward.

15. TOTAL RELEASES BY CARCINOGEN CLASSIFICATION :



The plot compares total chemical releases between substances classified as carcinogens ("YES") and non-carcinogens ("NO"). Both categories exhibit a wide range of release quantities, with most values clustered near the lower end and a few significant outliers with extremely high releases. The data suggests no immediate visible distinction in the overall distribution between carcinogenic and non-carcinogenic substances, though statistical analysis would be required to confirm differences. This visualization highlights the variability and potential environmental impact of chemical releases regardless of carcinogenic classification.

DATASET COMPARISON :

Before Exploration

```
print("Total number of rows and columns before exploration: ",data.shape)
data.head()
```

```
Total number of rows and columns before exploration:  (806162, 122)
```

	1. YEAR	2. TRIFD	3. FRS ID	4. FACILITY NAME	5. STREET ADDRESS	6. CITY	7. COUNTY	8. ST	9. ZIP	10. BIA	...	113. 8.2 – ENERGY RECOVERON	114. 8.3 – ENERGY RECOVEROF	115. 8.4 – RECYCLING ON SITE	116. 8.5 – RECYCLING OFF SIT	117. 8. – TREATMEN ON SIT
0	2018	28655MLDDP213RE	1.100004e+11	MOLDED FIBER GLASS NORTH CAROLINA	213 REEP DR	MORGANTON	BURKE	NC	28655	NaN	...	0.0	0.0	0.0	0.0	0.
1	2018	91744MNTXN13300	1.100005e+11	MAINTEX INC	13300 E NELSON AVE	CITY OF INDUSTRY	LOS ANGELES	CA	91746	NaN	...	0.0	0.0	0.0	0.0	0.
2	2018	15690SSVND130LI	1.100144e+11	ATI FLAT ROLLED PRODUCTS HOLDINGS LLC	130 LINCOLN AVE	VANDERGRIFT	WESTMORELAND	PA	15690	NaN	...	0.0	0.0	8535863.0	0.0	2735204.
3	2018	5320WCHRTR37WMI	1.100420e+11	CHARTER WIRE LLC	3700 W. MILWAUKEE ROAD	MILWAUKEE	MILWAUKEE	WI	53208	NaN	...	0.0	0.0	0.0	16.0	0.
4	2018	0808WPNDRL51SHA	1.100702e+11	PANDROL	501 SHARPTOWN RD	SWEDESBORO	GLOUCESTER	NJ	8085	NaN	...	0.0	0.0	0.0	3571.0	0.

5 rows × 122 columns

```
df1=pd.read_csv("/content/combined_tri_data_2014_2023.csv",low_memory=False)

null_counts = df1.isnull().sum()
null_counts = null_counts[null_counts > 0]  # Filter columns with at least one null value

# Display the counts
print("Count of columns with null values before exploration : ",null_counts.count())
```

```
Count of columns with null values before exploration :  25
```

After Exploration

```
(674021, 19)
```

	YEAR	FACILITY NAME	CITY	ST	INDUSTRY SECTOR	CHEMICAL	CARCINOGEN	WATER	OTHER DISPOSAL	ON-SITE RELEASE TOTAL	OFF-SITE RELEASE TOTAL	TOTAL TRANSFER
0	2017	DRS NIS LLC	DALLAS	TX	Computers and Electronic Products	Methanol	NO	0.0	0.0	11436.00	0.0	22113.00
1	2017	AMERICAN ELECTRIC POWER KAMMER/MITCHELL PLANT	MOUNDSVILLE	WV	Electric Utilities	Selenium compounds	NO	170.0	5.0	17200.00	2.0	2.00
2	2017	GEORGIA-PACIFIC LLC	TAYLORSVILLE	MS	Chemicals	Formaldehyde	YES	0.0	0.0	23640.00	5.0	948.00
3	2017	HITACHI ASTEMO AMERICAS INC. - BEREA KY	BEREA	KY	Transportation Equipment	Certain glycol ethers	NO	0.0	0.0	24674.45	0.0	7192.35
5	2017	SIGECO A B BROWN GENERATING STATION	MOUNT VERNON	IN	Electric Utilities	Polycyclic aromatic compounds	YES	0.0	0.0	1.00	0.0	0.00

```
null_counts = df.isnull().sum()
null_counts = null_counts[null_counts > 0]  # Filter columns with at least one null value

# Display the counts
print("Count of columns with null values after exploration : ",null_counts.count())
```

```
Count of columns with null values after exploration :  0
```

MODELS IMPLEMENTATION :

The objective was to forecast total chemical releases for the years 2024–2028 using multiple predictive models. Each model was evaluated based on its performance, specifically focusing on Root Mean Squared Error (RMSE). Below are the results of the models implemented, with forecasts shown in billions.

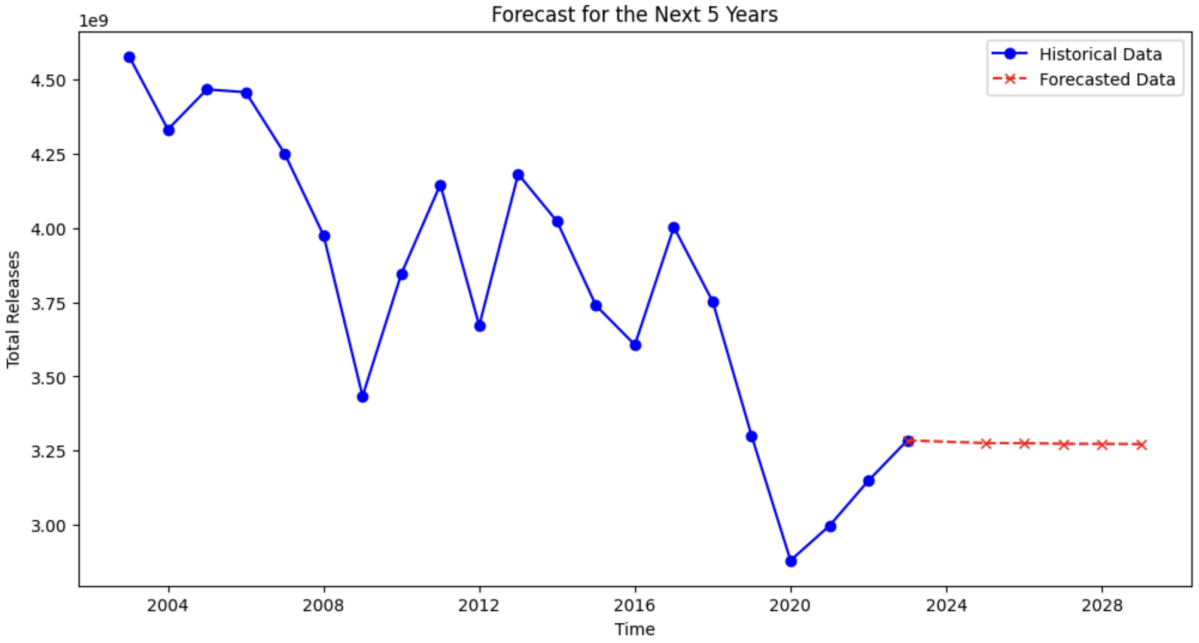
1. SARIMA (Seasonal AutoRegressive Integrated Moving Average)

Designed for time series data, SARIMAX captures trends, seasonality, and incorporates external factors, making it ideal for forecasting with seasonal patterns.

- **Best Parameters**: (2, 0, 2, 0, 0, 0, 12)
- **Best RMSE**: 1.12e8

Forecast :

YEAR	FORECASTED TOTAL RELEASE
2024	3,275,649,000
2025	3,275,362,000
2026	3,273,218,000
2027	3,272,890,000
2028	3,272,343,000



KEY INSIGHTS :

The forecast shows a consistent and minimal decrease in total chemical releases from 2024 to 2028, indicating stability in the industry with gradual improvements in emission control or chemical management.

2. ARIMA (AutoRegressive Integrated Moving Average)

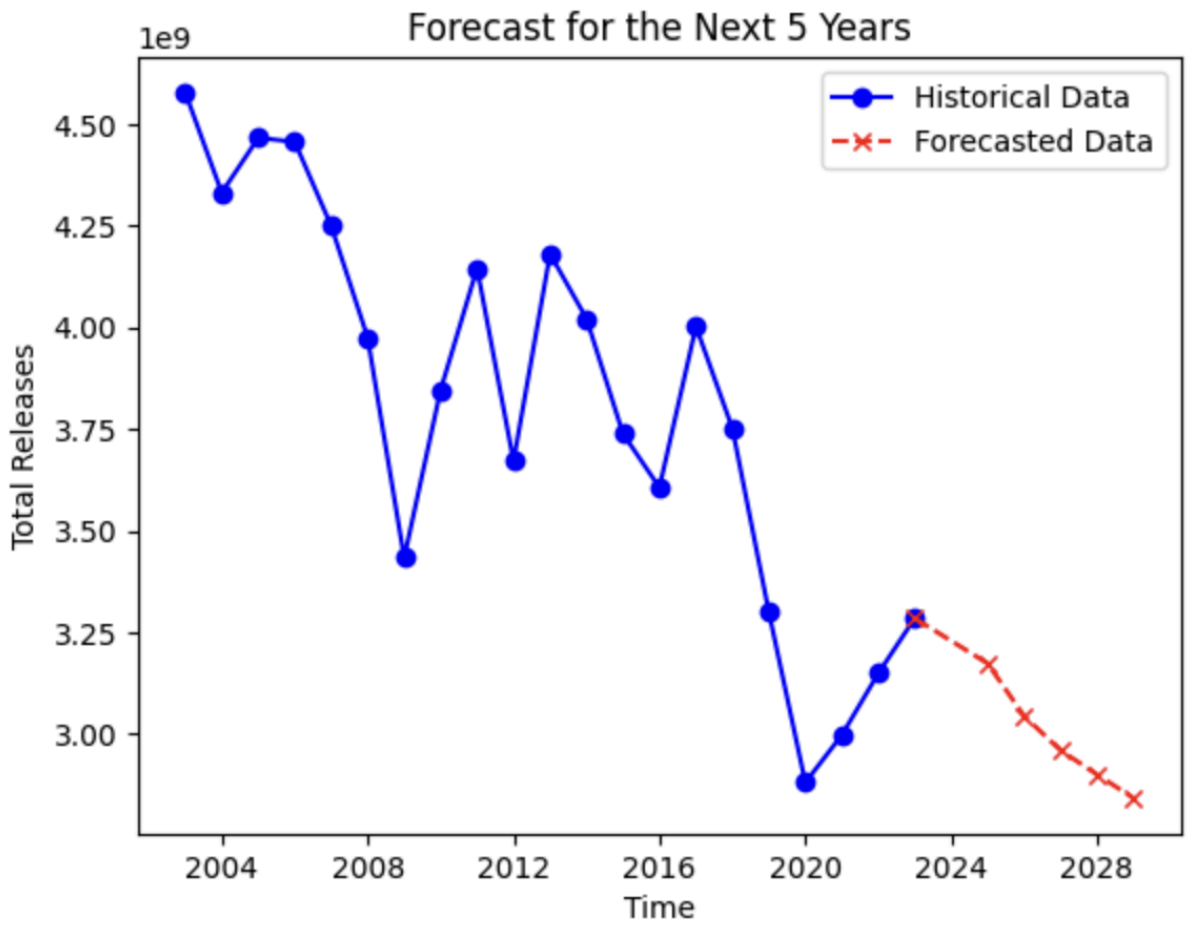
Best suited for stationary time series data, ARIMA models trends and patterns effectively, but it is limited to data without seasonal components.

- **Best Parameters**: (0, 0, 0)
- **Best RMSE**: 2.52e8

YEAR	FORECASTED TOTAL RELEASE
2024	3,173,345,000
2025	3,043,897,000
2026	2,959,638,000
2027	2,900,050,000

2028	2,841,968,000

Forecast :



KEY INSIGHTS :

The forecast indicates a steady decline in total chemical releases from 2024 to 2028, with a reduction of approximately 10% over the five-year period. This suggests improving efficiency or sustainability measures within the industry.

3. Prophet

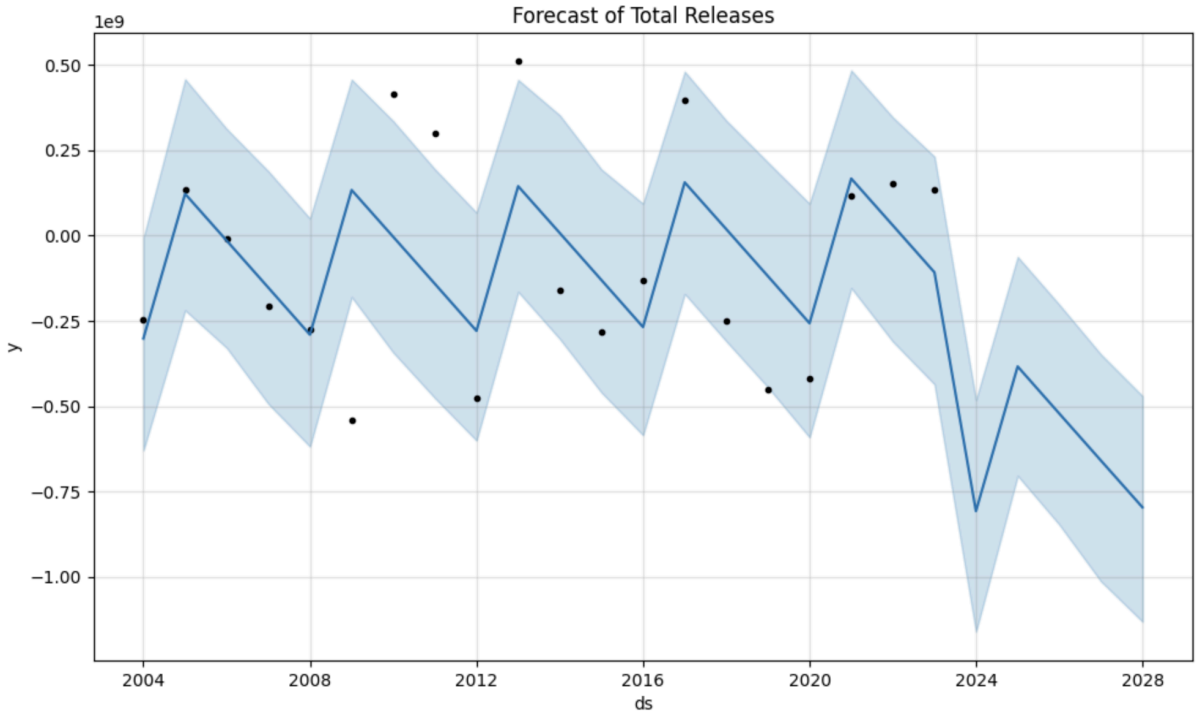
An intuitive forecasting tool designed for time series with trends and seasonality, Prophet is user-friendly but less suited for complex datasets.

- **Best RMSE**: 5.69e8

- **Parameters**: (0.5, 'additive')

Forecast :

YEAR	FORECASTED TOTAL RELEASE
2024	3,030,442,000
2025	2,967,577,000
2026	2,906,015,000
2027	2,845,731,000
2028	2,786,697,000



KEY INSIGHTS :

The forecast shows a gradual decrease in total chemical releases over the five years, with a steady decline from 2024 to 2028, highlighting ongoing efforts to reduce environmental impact.

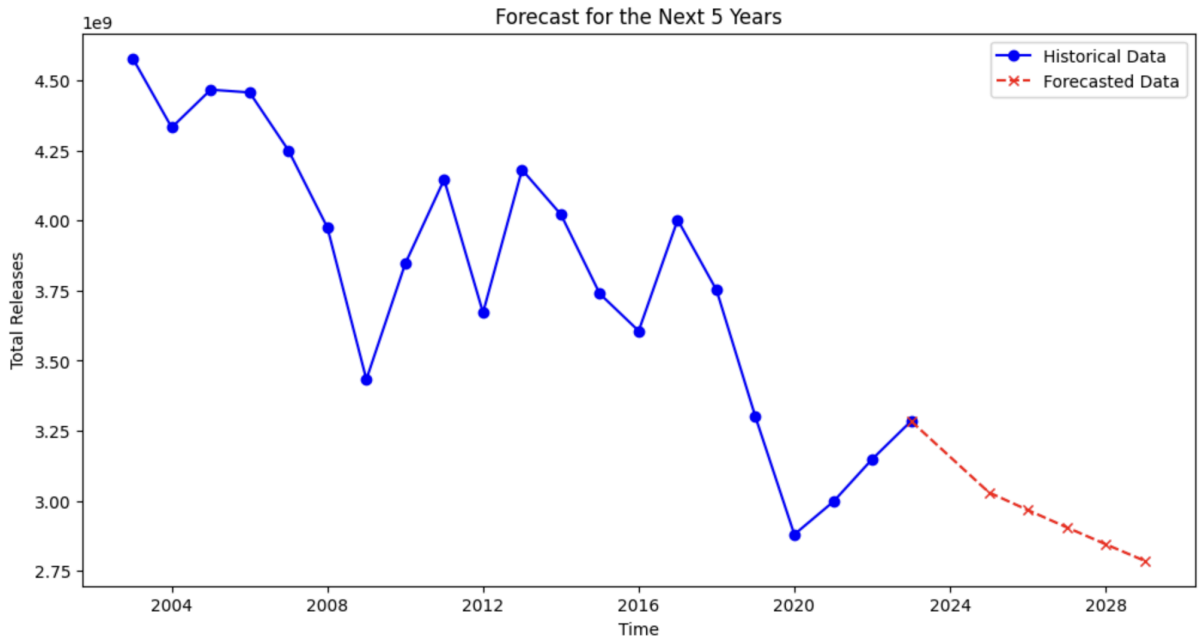
4. Exponential Smoothing (ETS - Holt-Winters)

This method uses weighted averages to forecast time series with seasonal trends, offering simplicity and reliability for straightforward datasets.

- **Best RMSE**: 4.10e8
- **Parameters**: ('multiplicative', None)

Forecast :

YEAR	FORECASTED TOTAL RELEASE
2024	3,030,442,000
2025	2,967,577,000
2026	2,906,015,000
2027	2,845,731,000
2028	2,786,697,000



KEY INSIGHTS :

The forecast shows a gradual decrease in total chemical releases over the five years, with a steady decline from 2024 to 2028, highlighting ongoing efforts to reduce environmental impact.

5. XGBRegressor (Extreme Gradient Boosting)

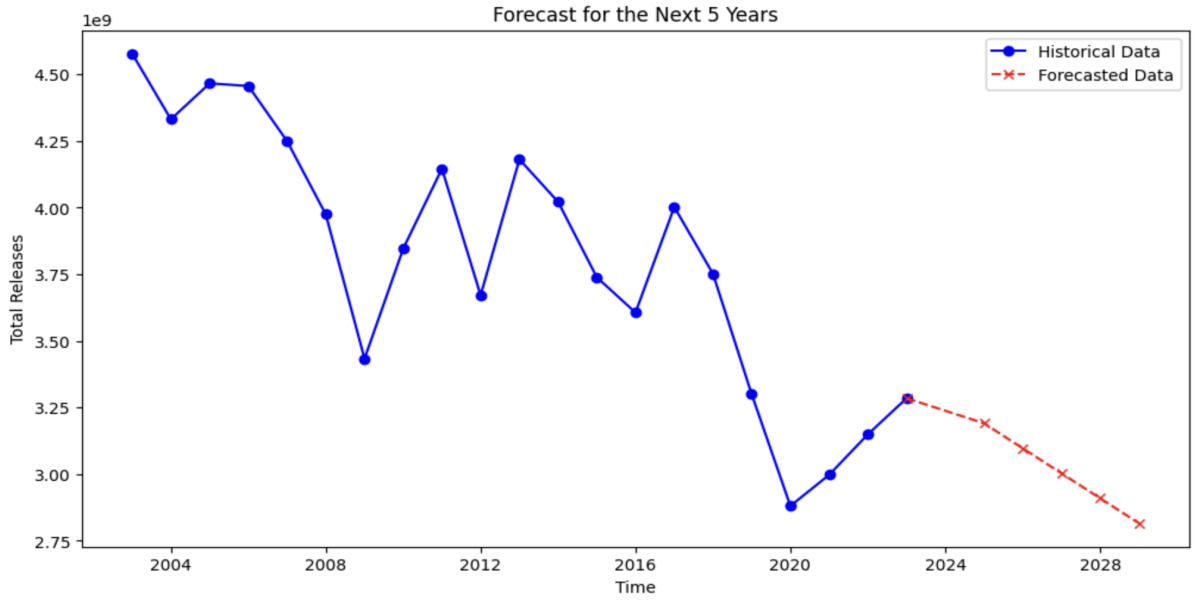
Known for its high accuracy, XGBoost is a robust model for capturing non-linear patterns in structured/tabular data.

- **Best RMSE**: 2.40e8

Forecast :

YEAR	FORECASTED TOTAL RELEASE
2024	3,190,811,000
2025	3,097,043,000
2026	3,003,274,000

2027	2,909,506,000
2028	2,815,738,000



KEY INSIGHTS :

The forecast shows a steady decrease in chemical releases from 2024 to 2028, suggesting ongoing improvements in chemical management or regulatory measures leading to a reduction in overall emissions.

6. LGBMRegressor (Light Gradient Boosting Machine)

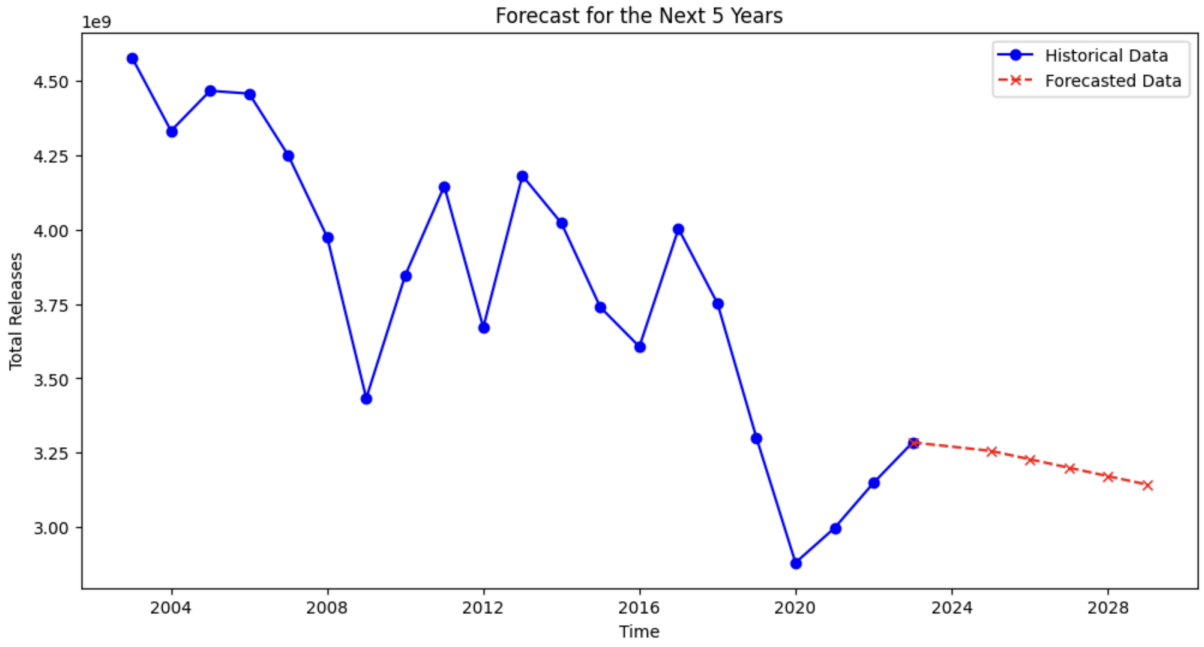
A fast and efficient machine learning model that excels in handling large datasets with complex relationships, making it a top choice for regression tasks.

- **Best RMSE**: 2.84e8

Forecast :

YEAR	FORECASTED TOTAL RELEASE
2024	3,256,304,000

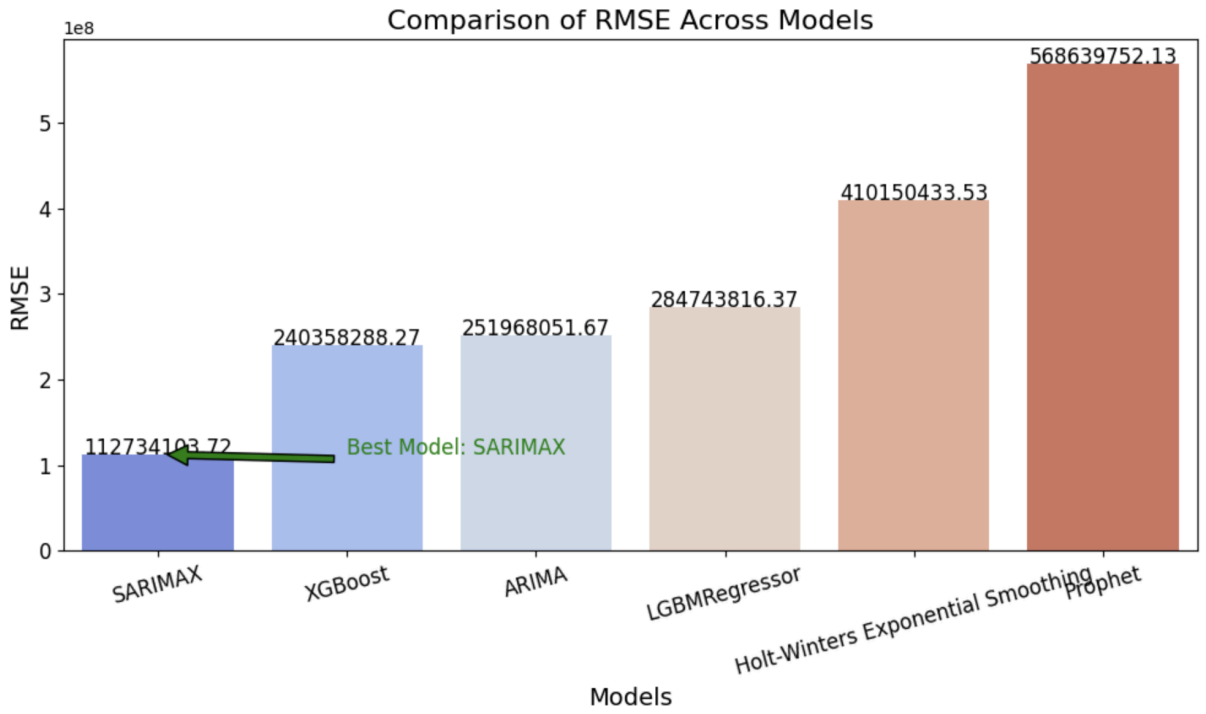
2025	3,228,029,000
2026	3,199,754,000
2027	3,171,479,000
2028	3,143,204,000



KEY INSIGHTS :

The forecast indicates a gradual decrease in chemical releases from 2024 to 2028, with a steady reduction over the years, pointing to incremental improvements in industry practices or compliance with environmental standards.

COMPARISON :



The chart showcases the **Root Mean Squared Error (RMSE)** for various forecasting models, measuring their prediction accuracy. A lower RMSE indicates a more accurate model.

RANKING (LOWEST RMSE TO HIGHEST RMSE):

- **SARIMAX (1.12e8)**: Best overall performance for time series data.
- **XGBoost (2.40e8)**: Second-best, suitable for non-linear relationships.
- **ARIMA (2.52e8)**: Good for stationary data without seasonality.
- **LGBMRegressor (2.84e8)**: Performed better than some statistical models but requires feature engineering.
- **Holt-Winters (4.10e8)**: Captures seasonality but less accurate for long-term forecasts.
- **Prophet (5.69e8)**: Easiest to use but least accurate in this scenario

KEY INSIGHTS :

- **Best Performing Model:**
    - SARIMAX is the most reliable, with the lowest RMSE of 1.12e8.
    - Forecasted values for total chemical releases show a slight downward trend after 2024, indicating stability in prediction.

- **Second Best Model:**

    - XGBRegressor demonstrated high performance with an RMSE of 2.40e8, suggesting it can effectively handle non-linear patterns.
- **Worst Performing Model:**
    - Prophet has the highest RMSE (5.69e8), making it the least accurate for this dataset.

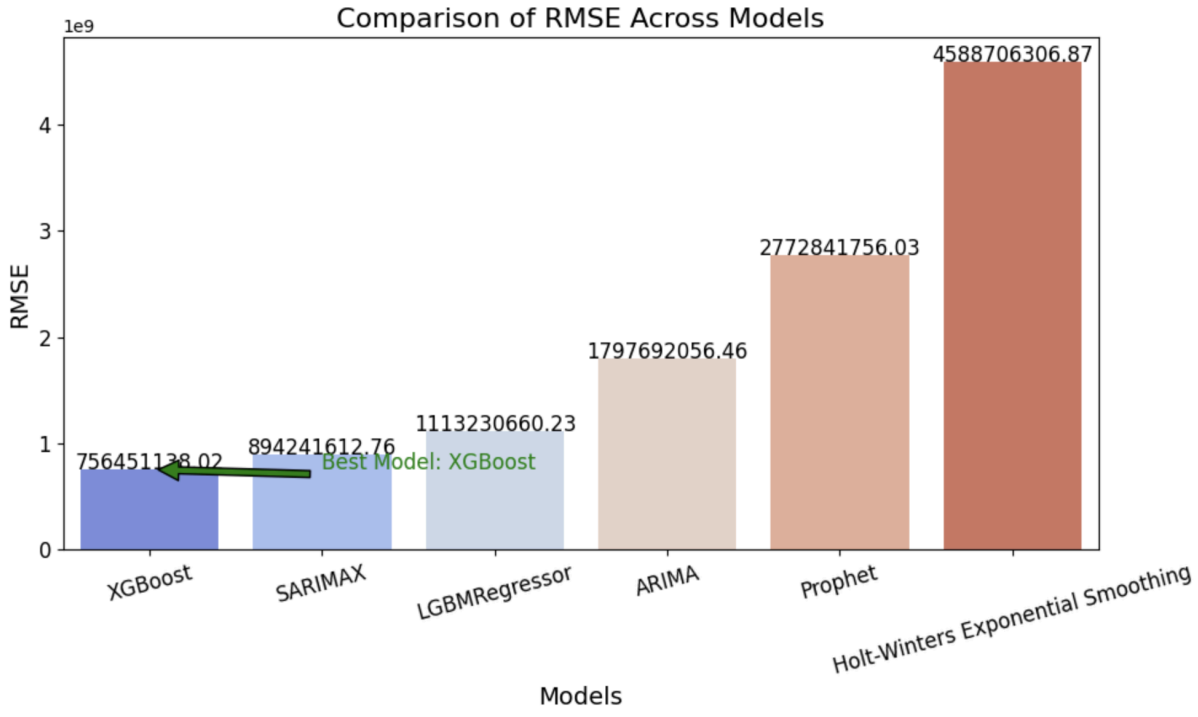
SARIMAX has the lowest RMSE (1.12e8), indicating it produced the most accurate predictions among the tested models.

EXTENDED MODEL IMPLEMENTATIONS :

In addition to forecasting total chemical releases, **similar methodologies were applied to predict other key metrics**, ensuring comprehensive analysis across various dimensions. These included:

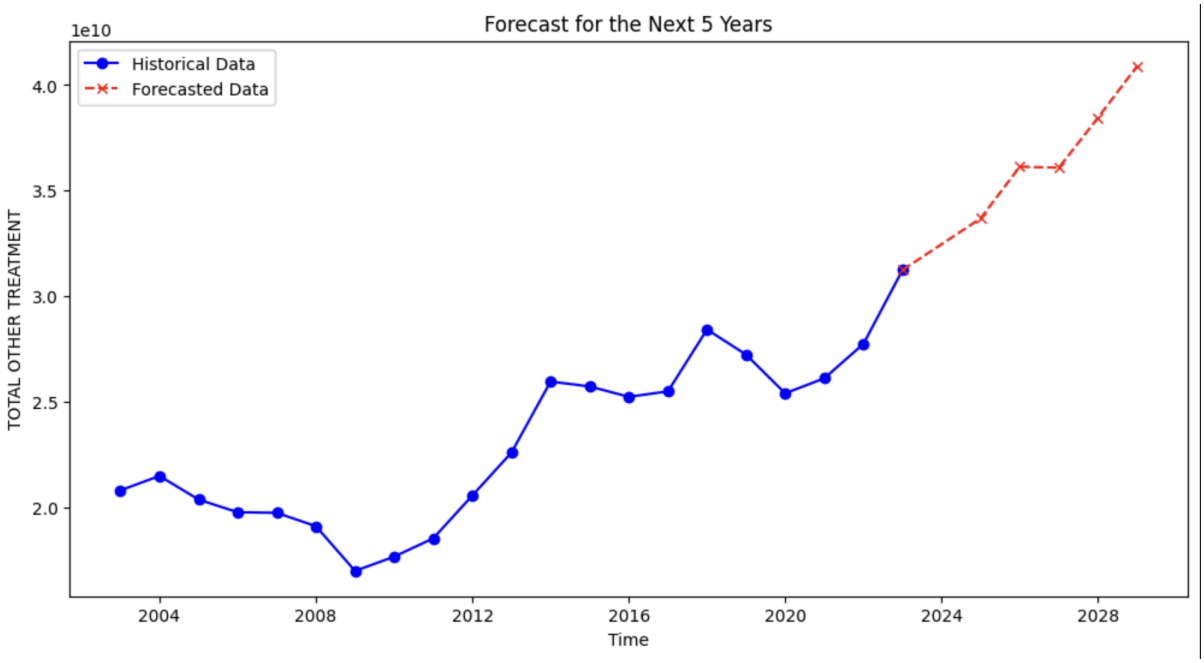
1. **Total Treatment (e.g., recycling and other non-direct releases):**
   Models were implemented to predict quantities treated through recycling, energy recovery, or other waste management techniques.

   BEST RMSE: 7.56e8



Forecast :

YEAR	FORECASTED TOTAL TREATMENT
2024	33,690,530,000
2025	36,134,980,000
2026	36,087,340,000
2027	38,449,500,000
2028	40,893,950,000



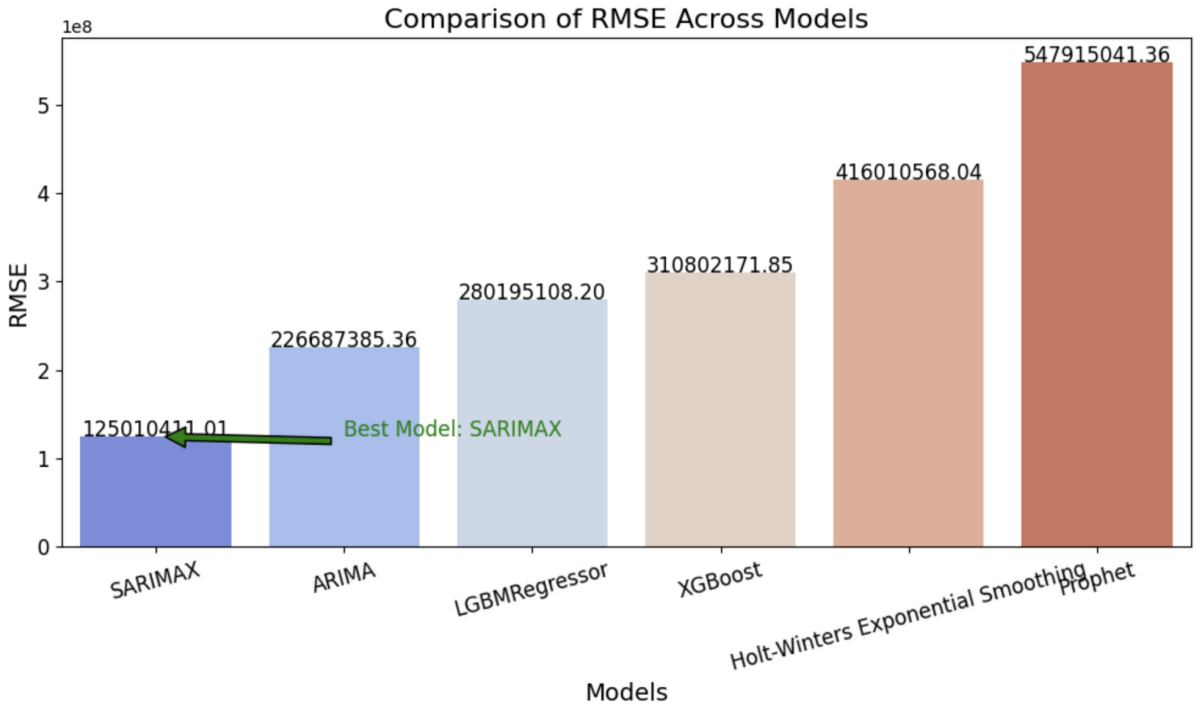
KEY INSIGHTS :

- The forecast shows a steady increase in total treatment across the years, with a notable rise in 2027 and 2028. This reflects increasing efforts toward chemical treatment and recycling.
- The model predicts a continued growth trajectory, indicating the need for expanding infrastructure and efficiency in managing treatment processes.

2. **On-Site Releases:**

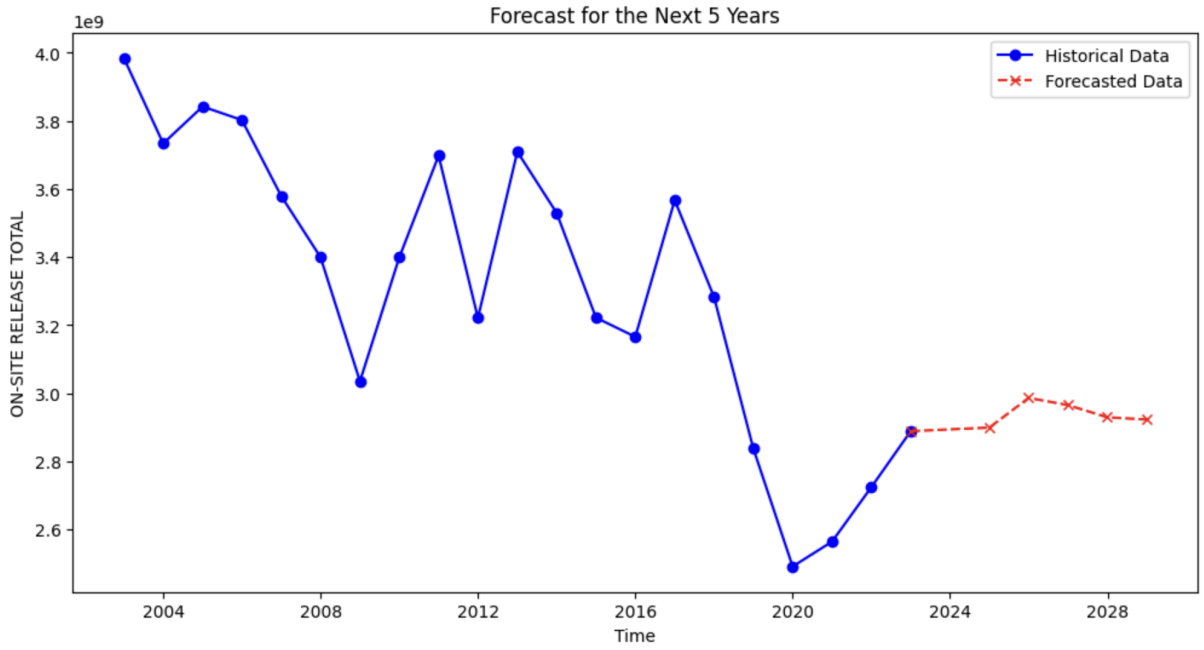
Focused on predicting chemical releases occurring within facility boundaries, providing insights into localized environmental impacts.

BEST RMSE: 1.25e8



Forecast :

YEAR	FORECASTED ON-SITE RELEASE
2024	2,899,630,000
2025	2,986,737,000
2026	2,965,778,000
2027	2,929,946,000
2028	2,923,376,000

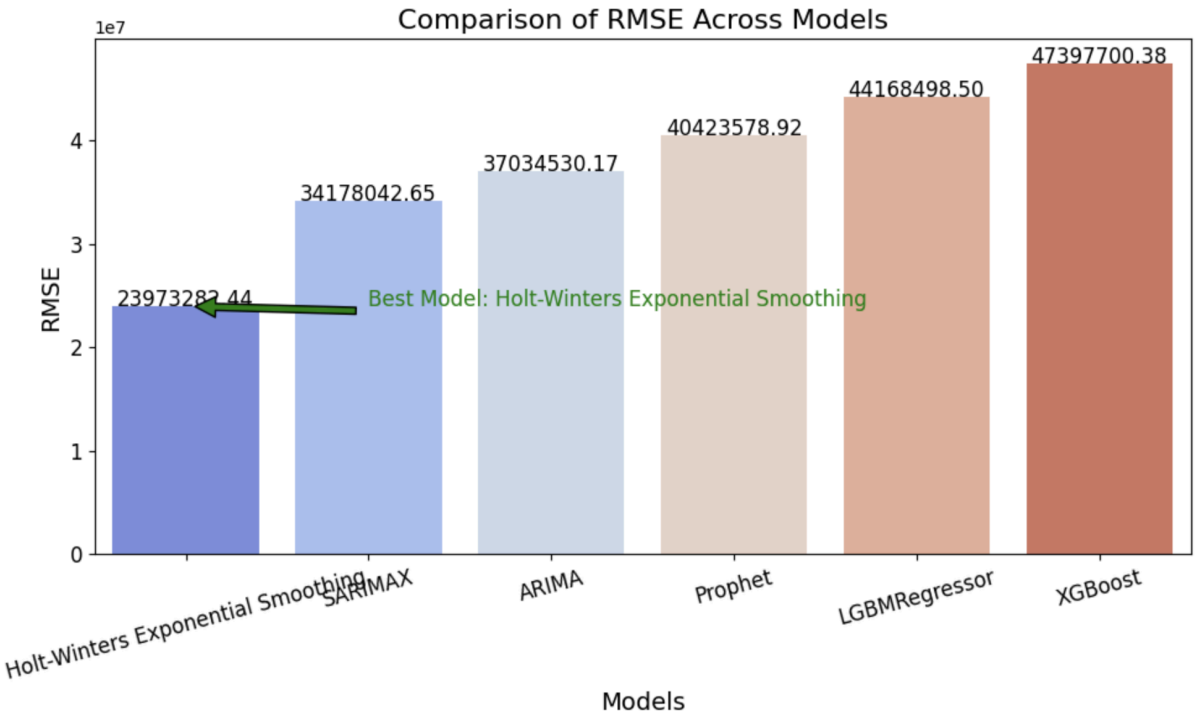


KEY INSIGHTS :

- The forecasted data indicates a relatively stable pattern for off-site releases. Although there is a slight increase in 2025, the following years show a marginal decrease in releases.
- The stability could suggest effective off-site chemical handling, but any fluctuations in the coming years should be closely monitored to ensure compliance with environmental standards.

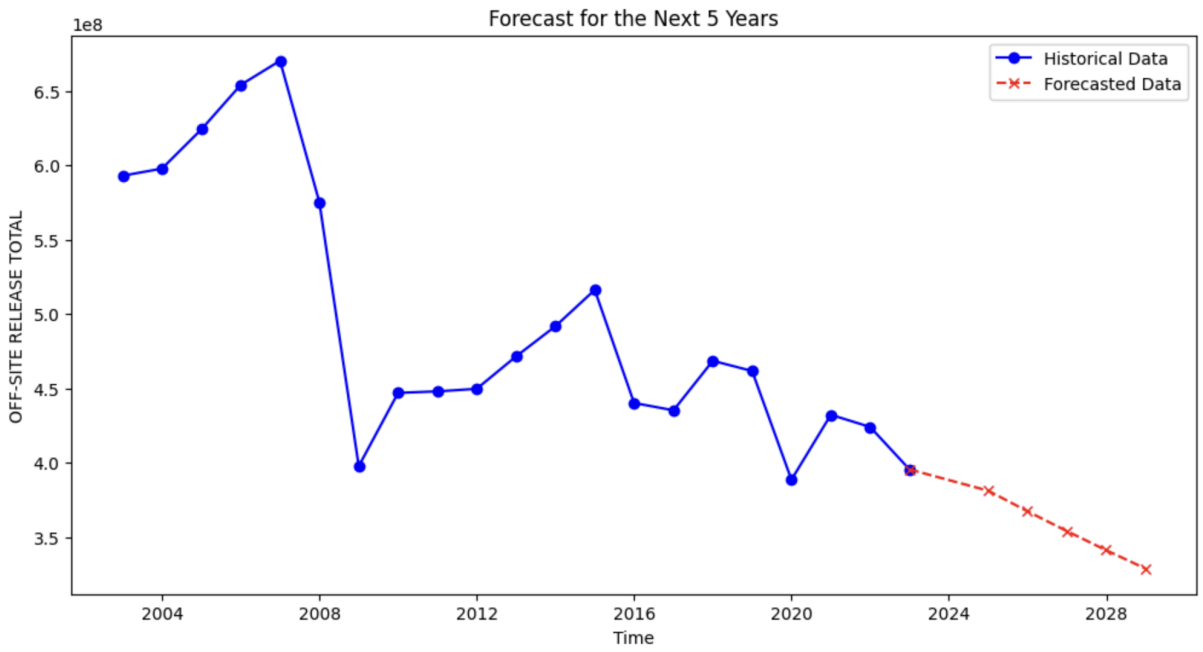
3. **Off-Site Releases:**
   Predicted quantities transported to off-site facilities for treatment or disposal, essential for understanding broader distribution.

   BEST RMSE: 2.40e8



Forecast :

YEAR	FORECASTED OFF-SITE RELEASE
2024	381,149,200
2025	367,309,000
2026	353,971,400
2027	341,118,100
2028	328,731,500



KEY INSIGHTS :

- The forecasted data indicates a gradual decline in on-site releases over the next 5 years. This could reflect improvements in chemical management practices, efficiency measures, or a reduction in chemical production at the facilities.
- Although the decline is expected, continuous monitoring and adjustments in regulations or operational processes may be necessary to maintain or improve this trend.

REFERENCES :

- **EPA Toxics Release Inventory (TRI) Program:** https://www.epa.gov/toxics-release-inventory-tri-program