Tox-e-mapper

Arya Kalappurayil, Cecili Poole, Diane Egret, Rhee Kang

1 Introduction

Environmental monitoring plays a crucial role in ensuring sustainable urban development and safeguarding public health. The release of toxic chemicals by industries can have serious environmental consequences, making it vital to monitor and manage these releases effectively. The Toxics Release Inventory (TRI)[2] provides a comprehensive database on the types and quantities of hazardous chemicals released into the environment by industries, offering transparency and a valuable tool for decision-makers. Existing platforms like EnviroMapper[1] lack user-friendly interfaces and effective tools for data exploration and analysis. The motivation behind this project is to improve the accessibility and usability of TRI data through an interactive geospatial UI, enabling stakeholders to make more informed decisions, track trends, and enhance pollution prevention efforts.

2 Problem Definition

Existing platforms that provide access to TRI data, such as EnviroMapper, lack the necessary functionality to enable effective analysis and interpretation of this information. The main challenge is the insufficient visualization of TRI data and the absence of interactive tools to track changes over time or evaluate the impact of regulations. This restricts stakeholders' capacity to effectively use the data for informed decision-making and policy assessment.

Our goal is to develop an interactive geospatial user interface (UI) that allows users to explore TRI data, clearly identify trends, and evaluate the effects of regulatory actions. This will enhance decision-making and environmental risk management for stakeholders.

3 Literature Survey

This literature survey explores existing research that informs the development of more effective tools for interpreting environmental data, focusing on issues of data accuracy, presentation, and usability.

Khanna (2019) [6] criticizes inconsistent production data and unreliable self-reporting by industries, especially regarding chemical releases. She advocates for removing "zero value" data, which influenced our data preparation by filtering out unnecessary information and enhancing data presentation for our geospatial platform. The TriSig paper [8] introduced a method for identifying significant patterns across three dimensions, which is used to highlight patterns linking pollution levels to specific times and places, inspiring our use of clustering techniques to pinpoint pollution hotspots.

The study on the Spatial Correlation of Industrial Carbon Emissions [9] highlighted regions with high emissions and their causes, offering insights for data-driven decision-making and recommending the region-specific policies for effective reduction of carbon footprints. This motivated us to recognize that focusing on regional levels is a crucial factor for stakeholders. Insights from the sales forecasting literature [16] guided our choice of metrics, where we opted for mean absolute error (MAE) for simplicity, instead of the suggested mean absolute scaled error (MASE). The comparative study on machine learning models for gas warning systems [13] informed our choice of Random Forests for their ability to handle complex data relationships, later deciding to test with histogram gradient boosting ensemble[14] to potential performance improvement.

The paper on data storytelling in visualization [3] emphasized the importance of narrative in improving data understanding, which guided the design of our geospatial data presentation. Techniques from Learning Geospatial Analysis with Python [7] helped process and visualize environmental data, despite limited coverage of environmental datasets like TRI. Additionally, studies using Folium for mapping and clustering, such as the Surabaya City criminal acts study [10] and the crime forecasting paper [11], inspired our approach for identifying pollution hotspots, although we faced challenges in clustering and integrating time-series data.

Finally, insights from Applied GeoSpatial DataScience with Python [5] and other papers on real-time data visualization and sensor data processing [12], as well as time series forecasting [4], contributed to shaping the tech stack for our interactive mapping and data handling. Additionally, scikit-learn's resources [15] helped guide the implementation of machine learning models,

particularly for evaluation metrics such as Mean Absolute Error(MAE), Mean Squared Error(MSE), Root Mean Squared Error(RMSE), and Mean Absolute Percentage Error(MAPE), which were critical in our experiments.

4 Proposed Method

The proposed methodology stands out from the state of the art by introducing a suite of innovative features designed to enhance both data analysis and decision-making processes. First, Dynamic Data Interactivity empowers users to explore TRI data through intuitive, interactive visualizations that support real-time filtering by location, chemical type, and time frame. This functionality enables stakeholders to engage more meaningfully with the data and make faster, better-informed decisions. Second, Advanced Clustering Analysis leverages techniques like PCA and K-Means to uncover pollution hotspots and guide effective pollution prevention strategies, offering a more analytical approach than traditional methods. Finally, Geospatial Visualization integrates facility-level maps with chemical release data, enhancing spatial awareness of environmental impacts and allowing for a clearer understanding of pollution distribution. An embedded dashboard further enhances user experience by consolidating key metrics and visual summaries in one accessible view. Together, these innovations intuitively provide richer insights, faster pattern recognition, and better decision support. The steps taken to develop tox-e-mapper are divided into five main sections as below:

(1) **Data Cleaning and Preparation**: Essential steps were taken to improve the TRI dataset's quality (1987–2023, 3 million data points):
   - Removed facilities with less than 15 years of data and applied an 80% null threshold, removing 15 fields with excessive missing data.
   - Removed 63 fields with over 80% zero values to focus on meaningful data.
   - Standardized Zip Code field by converting mixed data types (strings/numbers) to numeric format.

   These steps ensured a cleaner dataset, ready for analysis, modeling, and clustering.
(2) **Exploratory Data Analysis (EDA):** An in-depth EDA is conducted to understand patterns in the dataset[Refer Appendix 2], including the following steps:
   - Statistical Summaries: Computed descriptive statistics (mean, median, standard deviation, quartiles) for key variables like chemical releases, production waste, and facility locations to understand central tendencies and variability.
   - Feature Importance and Decision Tree Regressor: A Decision Tree Regressor was trained on 25% of the dataset, removing irrelevant fields. The top 10 important features were visualized using a color-coded bar chart to guide feature engineering and model selection.
   - Log-Scale Histogram of Key Features: Log-scale histograms for the top 5 features were created to examine skewed distributions, data patterns, and outliers.
   - Correlation Matrix and VIF: A correlation matrix heatmap was generated to detect multicollinearity, and the Variance Inflation Factor (VIF) was calculated to flag features for potential removal or transformation.
(3) **Model Training**: Machine learning models are used to analyze the data. The model training process involves the following steps:
   - Model Selection: Chose HistGradientBoostingRegressor for its ability to handle large datasets and predict continuous outcomes.
   - Hyperparameter Tuning: Conducted Grid Search to optimize parameters (learning_rate, max_leaf_nodes, regularization, loss_function) with four scoring functions.
   - Model Evaluation: Assessed performance using MAE, MSE, RMSE, MAPE across three runs; analyzed learning curves for overfitting/underfitting[Refer Appendix 1].
   - Error Handling: Monitored and adjusted the model's tendency to predict zeros (no chemical release) for meaningful predictions.
(4) **Clustering**
   - PCA for Dimensionality Reduction: Principal Component Analysis (PCA) is used to reduce the dataset's dimensionality, preserving variance and improving clustering performance by focusing on the most significant features.
   - K-Means Clustering: The K-Means algorithm groups facilities with similar chemical release profiles. The optimal number of clusters, K, is determined using the elbow method and silhouette analysis.

- Cluster Interpretation: Clusters are analyzed to identify common characteristics (e.g., chemical releases, industry types), which inform pollution prevention strategies and regulatory decisions. These clusters are visualized on a geospatial map.

(5) **User Interface Building**:

- **Data Visualization and Interactive Features:** Tableau visualizations[Refer Figure 1 Appendix 3] and geospatial mapping allow users to filter TRI data by location, chemical, industry, and time, with mouse hover details on facility locations. User-friendly filters, selection cards, and clustering models enable refined searches and deeper insights.
- **Embedding and Deployment:** The final dashboard is embedded in a website using JavaScript, HTML, and CSS for easy access, promoting public interaction and data-driven decision-making.

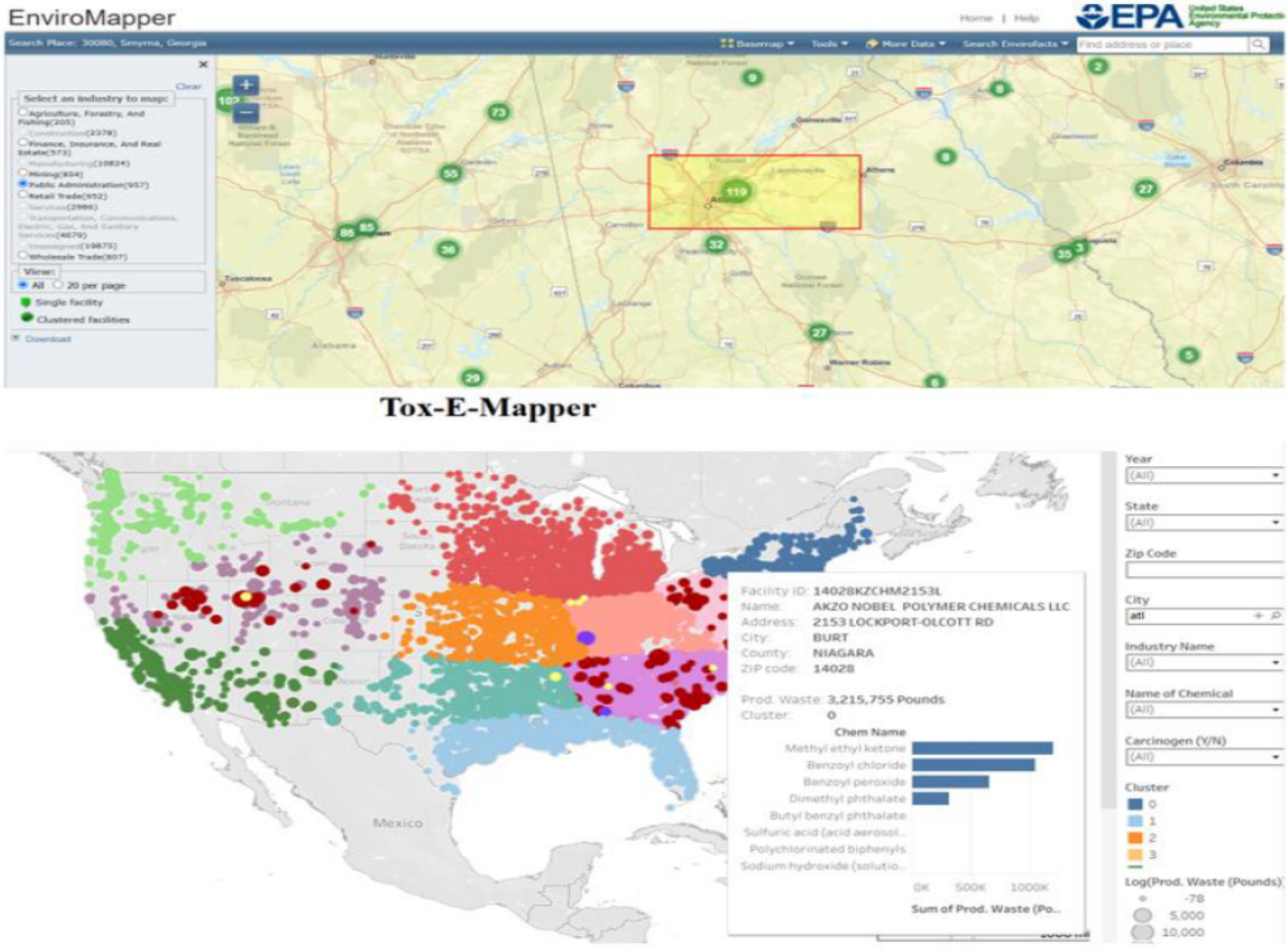


Figure 1: Enviromapper Vs Tox-e-mapper

5 Evaluation

The evaluation of Tox-E-Mapper focuses on assessing the platform's usability and its effectiveness in supporting data exploration and pattern recognition. The evaluation involved participants including classmates and general environmental data users. The goal was to understand how Tox-E-Mapper performs compared to the EnviroMapper tool, particularly in terms of user

experience and data exploration capabilities. The evaluation is designed to answer the following research questions:

(1) Is Tox-E-Mapper more user-friendly and interactive compared to EPA's EnviroMapper?
(2) Does Tox-E-Mapper reduce task completion time and improve user satisfaction?
(3) How easy is it to identify pollution hotspots and explore chemical release trends using Tox-E-Mapper?

5.1 Experimental Design and Observations

- **Usability Evaluation:**To assess the usability of Tox-E-Mapper and EnviroMapper, a user survey was conducted in which participants completed a series of common environmental data-related tasks using both platforms. The tasks included locating the user's address to view TRI data, identifying the most recently released toxin in their area, and exploring the emission trend of a selected toxin over time. Participants rated how easy or difficult it was to complete each task on a 5-point Likert scale, where a rating of 1 indicated "Very Easy" and 5 indicated "Very Difficult." The primary goal of the survey was to gauge user satisfaction, ease of use, and task efficiency for both tools.
  The results clearly demonstrated that Tox-E-Mapper was significantly more user-friendly than EnviroMapper. Tox-E-Mapper had a mean ease of use rating of 1.79, while EnviroMapper scored 3.29, indicating that participants found Tox-E-Mapper considerably easier to use. For specific tasks, Tox-E-Mapper was notably faster and easier to navigate, with ratings of 2.00 for Task 1 (locating TRI data) and 2.43 for Task 2 (identifying the most recently released toxin), compared to 3.71 for both tasks in EnviroMapper. These findings suggest that Tox-E-Mapper not only facilitated quicker task completion but also offered a more satisfying and efficient user experience overall.
- **Data Exploration and Pattern Recognition:**The second part of the evaluation focused on how well participants were able to explore environmental data, identify pollution hotspots, and analyze chemical release trends using Tox-E-Mapper's visual tools. Participants were tasked with identifying high-concentration areas of chemical emissions, interpreting trends in toxin releases over time, and exploring the broader environmental context by filtering data by chemical type and timeframe.

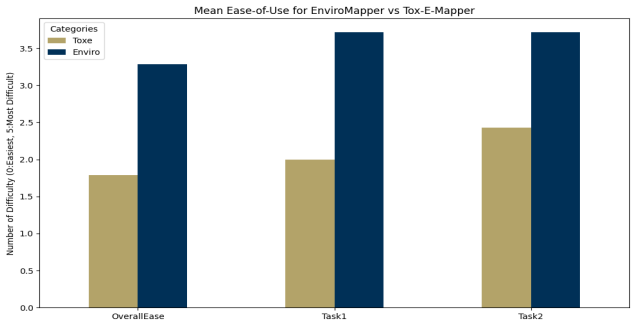


Figure 2: Mean ease-of-use for Enviromapper Vs Tox-e-mapper

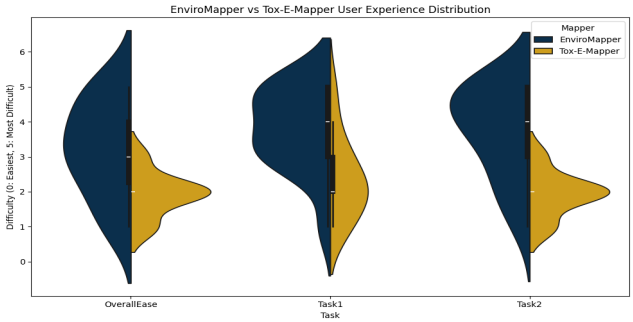


Figure 3: User experience distribution: Enviromapper Vs Tox-e-mapper

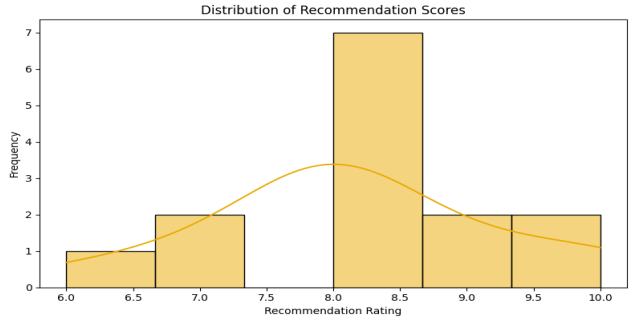


Figure 4: Distribution of recommendation score for Tox-e-mapper

Participants found the clustering feature in Tox-E-Mapper particularly useful for identifying pollution hotspots. The map view, enhanced by clustering overlays, made it easier to spot regions with the highest emissions. In contrast, EnviroMapper's map lacked the same level of interactivity and data density, making it more difficult for users to quickly identify these hotspots.

When analyzing chemical emission trends, participants were able to easily interpret whether toxin levels were increasing or decreasing over time. The ability to filter data by chemical type and timeframe allowed users to tailor their analysis to the most relevant data, providing a clearer understanding of emission trends in their specific region. This capability enhanced the overall exploration experience, making it more intuitive and efficient.

These findings confirm that Tox-E-Mapper outperforms EnviroMapper in supporting users in exploring environmental data and identifying key patterns in chemical release, thanks to its more advanced visualization tools and enhanced interactivity.

6 Conclusions and Discussions

In this project, we developed Tox-E-Mapper, an interactive geospatial tool aimed at improving the accessibility and usability of the Toxics Release Inventory (TRI) data. Our primary goal was to enhance the ability of stakeholders to monitor, analyze, and make data-driven decisions regarding hazardous chemical releases. By leveraging advanced clustering techniques, dynamic data interactivity, and geospatial visualizations, Tox-E-Mapper enables users to track environmental trends more effectively than traditional platforms like EnviroMapper.

6.1 Key Findings and Results

(1) **User Experience:** Tox-E-Mapper outperforms EnviroMapper in ease of use, task completion speed, and user satisfaction. Participants found it faster and more intuitive for identifying pollution hotspots and analyzing emission trends.

(2) **Data Analysis and Clustering:**The use of PCA and K-Means clustering effectively identified pollution hotspots, offering better insights into chemical release patterns than traditional tools.

(3) **Machine Learning Models:**The machine learning models employed, particularly the HistGradientBoostingRegressor, demonstrated good potential in analyzing the TRI dataset.However, issues with predicting zeros (no chemical releases) limited the models' ability to offer valuable insights for forecasting. Despite these challenges, the models successfully highlighted important features and provided a robust foundation for the platform's data exploration tools.

Scoring Function	*MAPE*	*MAE*	*MSE*	*RMSE*
neg_mean_absolute_percentage_error	0.26	118481.36	54316723669698.88	7369988.04
neg_mean_squared_error	1.37e+20	143174.61	53606213011906.93	7321626.39
neg_root_mean_squared_error	0.26	118481.36	54316723669698.88	7369988.04
neg_mean_absolute_error	0.26	118481.36	54316723669698.88	7369988.04

Table 1: Sample Metrics from 1 experiment training model for forecasting with Histogram Gradient Boosting Regressor

6.2 Limitations

(1) **Model Performance and Limitations:** During model training, evaluation scores and learning curves indicated that the model was learning. However, it frequently predicted zeros, limiting its ability to provide meaningful insights. This led us to drop our proposed forecasting innovation, as more data is needed to improve model performance.

(2) **Tool and Database Choices:** Initially, we considered using D3.js or Folium for geospatial visualizations, but we ultimately switched to Tableau for

its practicality, ease of use, and superior capabilities in creating interactive, dynamic visualizations that better aligned with the project's needs. Similarly, while we had planned to use PostgreSQL for database management, the high data volume proved costly, prompting us to opt for SQLite, which provided a more cost-effective solution for handling large datasets.

(3) **Incomplete Dataset:** The dataset only includes government-reported industries, excluding data from non-reporting facilities, which limits the comprehensiveness of the analysis.

6.3 Future Scope

(1) **Forecasting and Alternative Approaches:**Explore classification or anomaly detection methods for better handling of data sparsity and improve model performance.

(2) **Platform Accessibility:**Host Tox-E-Mapper publicly by purchasing a domain and expanding accessibility.

In conclusion, Tox-E-Mapper improves TRI data analysis through interactive geospatial visualizations, clustering, and dynamic filtering. While model performance was limited by frequent zero predictions, preventing meaningful insights for forecasting, the platform still offers valuable tools for environmental decision-making. The project highlights the potential of combining machine learning and geospatial analysis for enhanced environmental monitoring, though more data is needed to refine the model. All team members contributed equally to the development and success of the project, collaborating on various aspects such as data cleaning, machine learning, UI design, and usability evaluation.

References

[1] EPA. Enviromapper. https://enviro.epa.gov/envirofacts/enviromapper/search, 2025.

[2] EPA. Toxics release inventory (tri) program. https://www.epa.gov/toxics-release-inventory-tri-program, 2025.

[3] Vanessa Echeverria Lixiang Yan Dragan Gasevic Hongbo Shao, Roberto Martinez-Maldonado. Data storytelling in data visualisation: Does it enhance the efficiency and effectiveness of information retrieval and insights comprehension, 2024. Literature Survey by Cecili A Poole.

[4] Preeti Pandey Jitendra Singh. Metrics and scoring: quantifying the quality of predictions, 2024. Literature Survey by Arya Divakaran Kalappurayil.

[5] David S. Jordan. *Applied Geospatial Data Science with Python.* O'Reilly Media, 2023. Literature Survey by Arya Divakaran Kalappurayil.

[6] A. Khanna. Cornell dyson working paper (khanna, 2019), 2019. Literature Survey by Rhee J Kang.

[7] Joel Lawhead. *Learning Geospatial Analysis with Python.* O'Reilly Media, 2023. Literature Survey by Diane A Egret.

[8] Rui Henriques Leonardo Alexandre, Rafael S. Costa. Trisig: Assessing the statistical significance of triclusters, 2023. Literature Survey by Rhee J Kang.

[9] Rui Henriques Leonardo Alexandre, Rafael S. Costa. Study on spatial correlation and driving factors of industrial carbon emissions in china, 2025. Literature Survey by Rhee J Kang.

[10] Reisa Permatasari; Dhian Satria Yudha Kartika; Muhammad Daffa; Abdul Rezha Efrat Najaf; Bonda Sisephaputra; Nur Lukman. Unveiling patterns: Utilizing folium for visualizing clustered criminal acts distribution in surabaya city, 2023. Literature Survey by Diane A Egret.

[11] Akshay Jawla Nishtha Hooda, Manjot Singh. Crime forecasting using folium, 2020. Literature Survey by Diane A Egret.

[12] Vilem; Hejlova Vendula. Pohanka, Tomas; Pechanec. Python web server for sensor data visualization, 2016. Literature Survey by Arya Divakaran Kalappurayil.

[13] Huan Zhang Haiyan Lu Ergun Gide Jinrong Liu Clement Franck Benoit Charbonnier Robert M. X. Wu, Niusha Shafiabady. Comparative study of ten machine learning algorithms for short-term forecasting in gas warning systems, 2024. Literature Survey by Cecili A Poole.

[14] scikit learn. https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_hist_grad_boosting_comparison.html#sphx-glr-auto-examples-ensemble-plot-forest-hist-grad-boosting-comparison-py.

[15] scikit learn. https://scikit-learn.org/stable/modules/model_evaluation.html#mean-absolute-percentage-error.

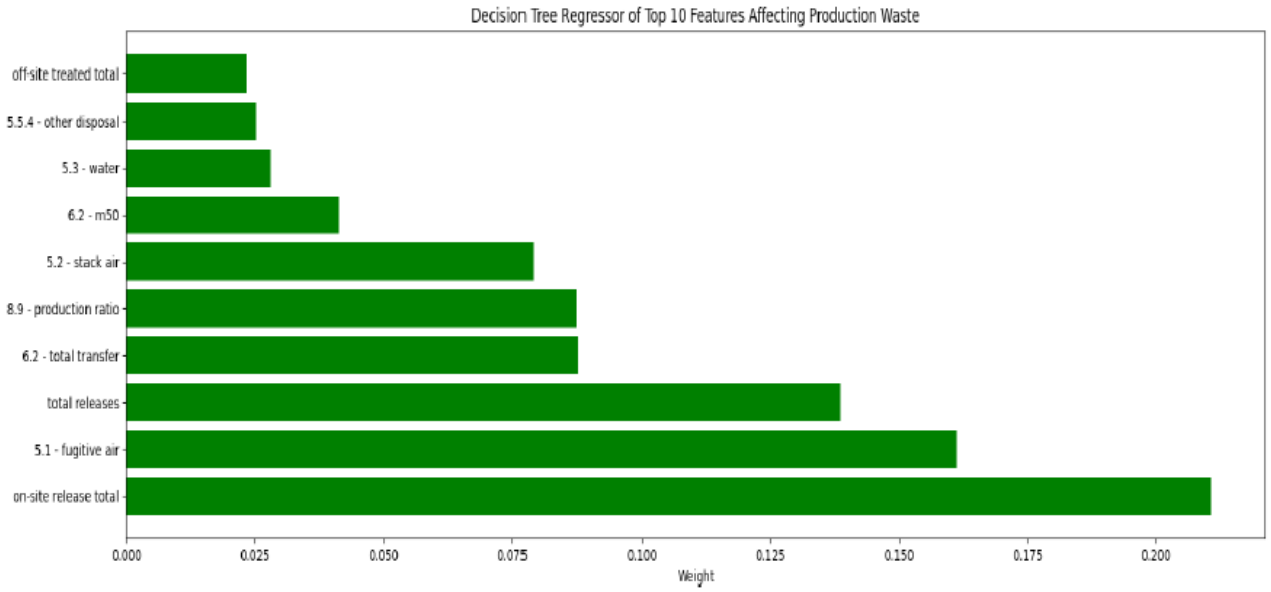
[16] Ralf W. Seifert a c Yara Kayyali Elalem a, Sebastian Maier b a. A machine learning-based framework for forecasting sales of new products with short life cycles using deep neural networks, 2023. Literature Survey by Cecili A Poole.

A Appendix 1 - Sample Metrics

Table 2: Sample Metrics from 1 experiment training model for forecasting with Histogram Gradient Boosting Regressor

Scoring Function	***MAPE***	***MAE***	***MSE***	***RMSE***
neg_mean_absolute_percentage_error	0.26	118481.36	54316723669698.88	7369988.04
neg_mean_squared_error	1.37e+20	143174.61	53606213011906.93	7321626.39
neg_root_mean_squared_error	0.26	118481.36	54316723669698.88	7369988.04
neg_mean_absolute_error	0.26	118481.36	54316723669698.88	7369988.04

B Appendix 2 - EDA Results



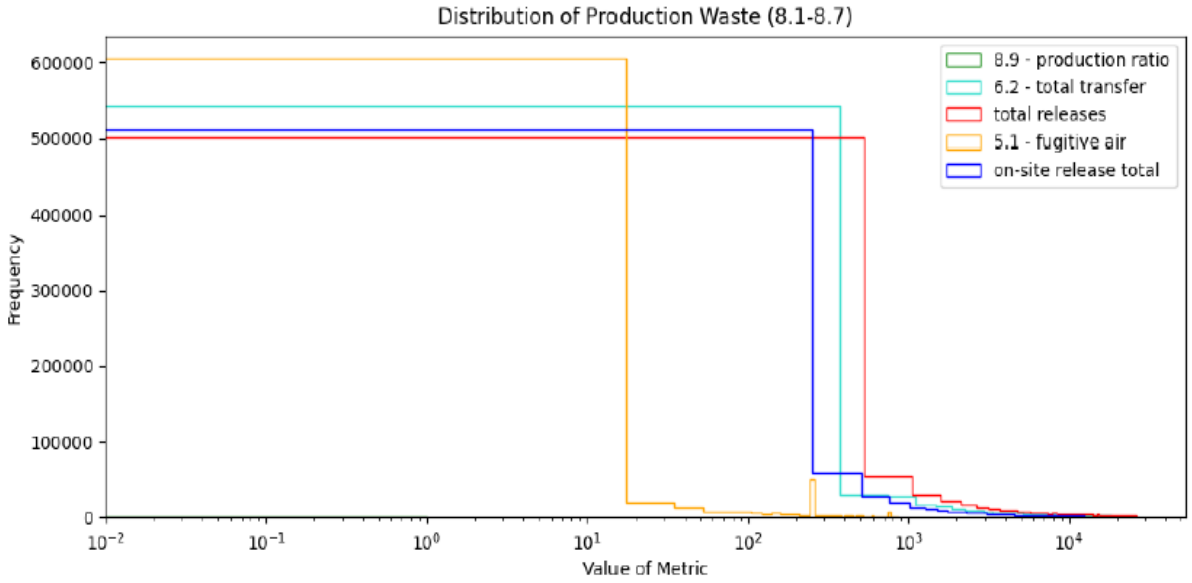


Figure 5: EDA visual output for DecisionTreeRegressor Production Waste Distribution

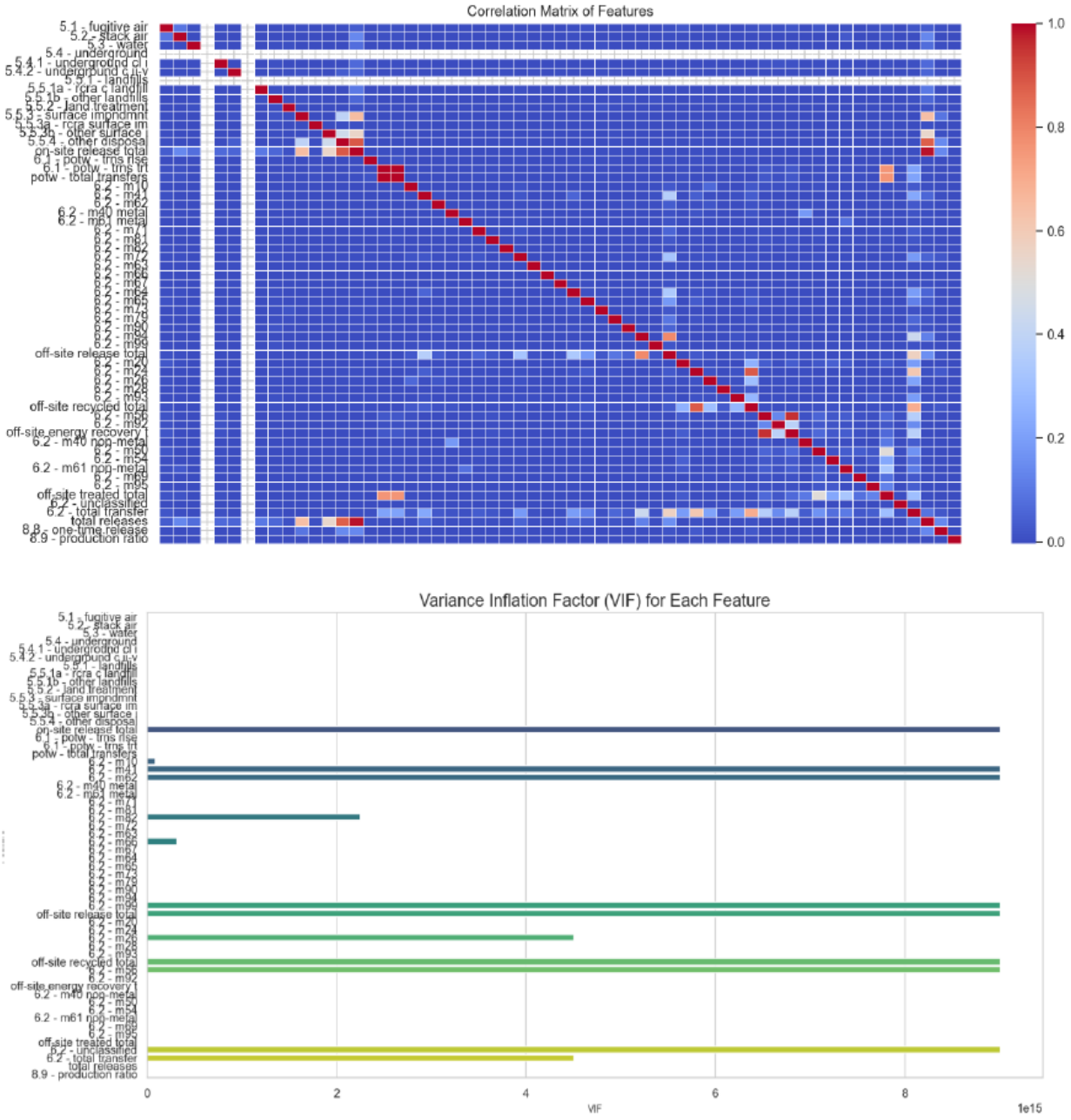


Figure 6: EDA visualization for correlation matrix VIF for each feature

C Appendix 3 - Tableau Prep Workflow

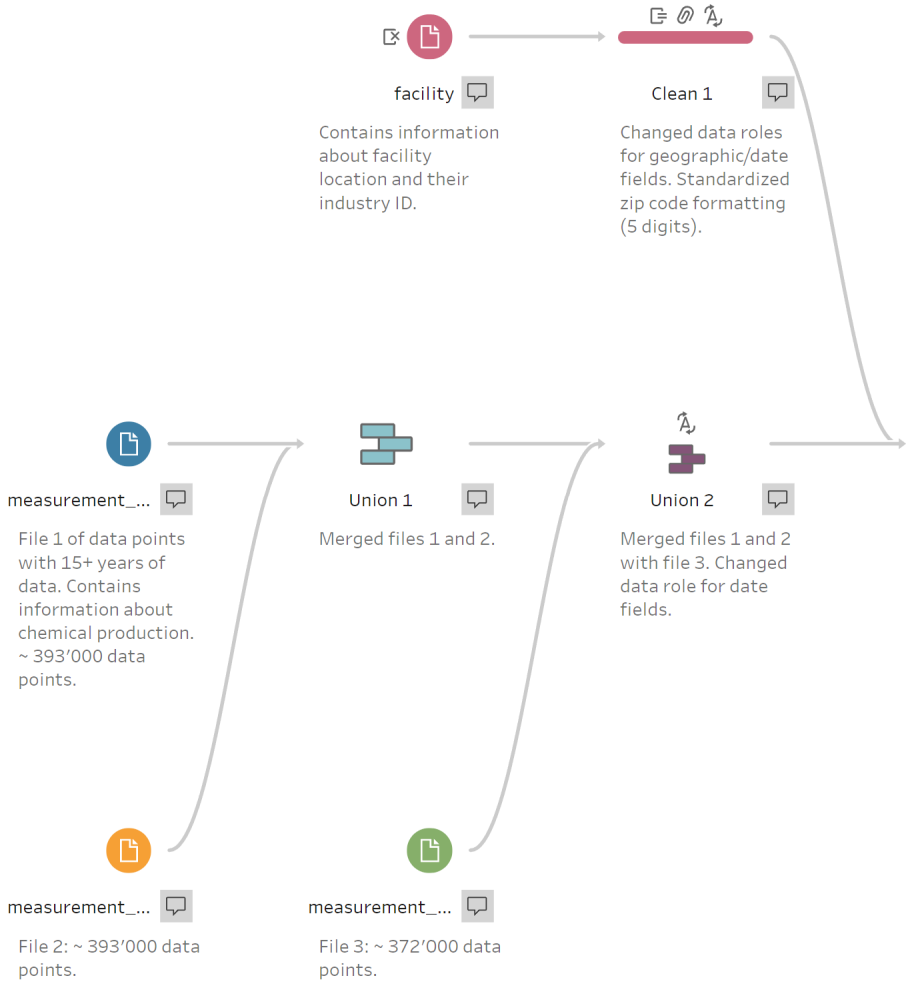


Figure 7: Tableau Prep Workflow - 1

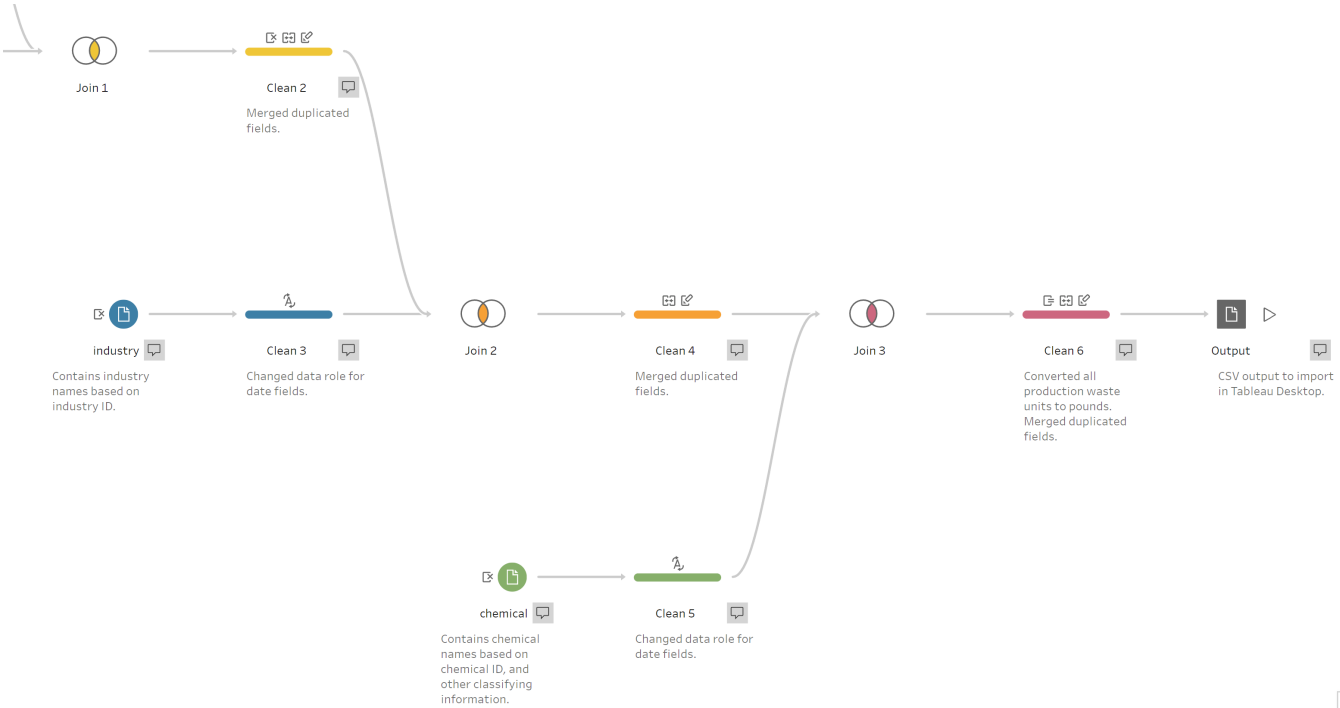


Figure 8: Tableau Prep Workflow - 2