Northeastern University

Final Project

Submission

Course: ALY6015

Intermediate Analytics

Prof. Wada Roy

Submitted

By

Hari Babu Muga

Ajay Reddy Potlapally

Akanksha Mamidi

Harsha Vardhan Reddy Mogulla

Sec-01

19/02/2022

Introduction

This assignment is the first step towards the final project for this course. We are 4 students who have joined group 7 and decided to work on the Toxic Release Inventory dataset for our final project. As part of this assignment, we would like to present our findings of the dataset, questions we want solve, approached to follow for building machine learning models for predictions.

Dataset

The Toxic Release Inventory tracks the management of toxic chemicals that may pose a threat to human health and the environment. This dataset records annual volume of toxic chemicals disposed and managed by almost 22,000 facilities in the US (United States) from 1987 – 2014. These data also include the total release of 30 most common chemicals tracked by this program, total release of metal and carcinogen for each company/facility.

Title: Threat Prediction to Human Health and Environment

Questions

1. Find the highest amount of chemicals released each year, and predict its trend?
2. Which facility releases the highest amount of toxic compounds in total each year?
3. Which compounds have the highest correlation?
4. Which compounds have the lowest correlation?
5. Find the Total Release of top 30 Chemicals in all Years?
6. Find the Total Release of metal and Carcinogen Chemicals?
7. Find the Highest number of Chemicals Released by each state?
8. Find the Lowest number of Chemicals Released by each state?
9. Does every county have the same threat?
10. What are the most affected cities because of the excessive chemical releases?
11. Is there any seasonal pattern behavior associated with any company?
12. What are the causes of excessive chemical releases and how to prevent them?
13. How to build the best pipeline using analytical methods for identifying the threats and prevent them early?

Analytical Methods and Plans

As of now, with the skillset gained from the Introduction to Analytics course and the analytical methods that we are going to learn in this Intermediate Analytics course, we would like to build a pipeline of below mentioned analytical methods to find answers to the above listed questions.

1) Descriptive Statistics

2) Data Imputation

3) Exploratory Analysis

4) Correlation Analysis

5) Hypothesis Testing

6) Data Scaling & Encoding

7) Data Splitting

8) Regression Analysis

9) Models Building

10) Models Evaluation

11) Predictive Analytics

12) Time Series Analysis

1) Descriptive Statistics

Given dataset has a total of 680557 samples and 49 attributes. Missing values are found in the dataset which we will impute using measures of central tendency statistics. Below are the attribute data types and their counts.

Data Type	Count
Float64	40
Int64	1
Object	8

Let us look at the statistics of a few continuous attributes below.

Statistics	ammonia	benzoghiperylene	total_release_carcinogen	total_release_metal
count	6.805570e+05	394678.000000	6.599030e+05	6.599030e+05
mean	9.836033e+03	15.039284	1.124503e+04	1.211721e+05
std	2.529126e+05	515.280164	1.413028e+05	4.465320e+06
min	0.000000e+00	0.000000	0.000000e+00	0.000000e+00
25%	0.000000e+00	0.000000	0.000000e+00	0.000000e+00
50%	0.000000e+00	0.000000	0.000000e+00	0.000000e+00
75%	0.000000e+00	0.000000	2.263850e+02	4.400000e+02
max	5.810000e+07	65112.801000	4.030000e+07	1.110000e+09

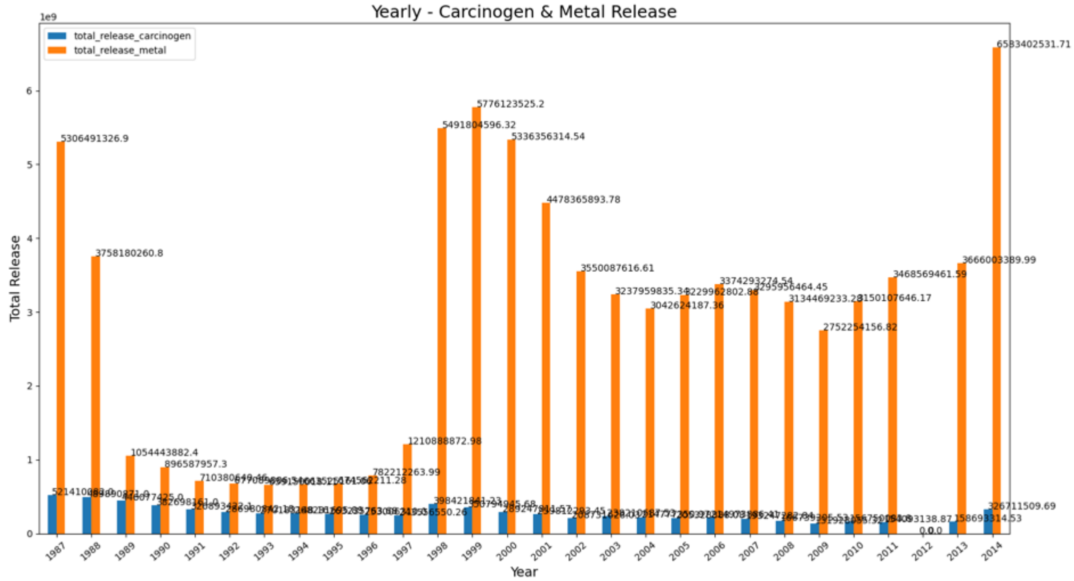
It is clear from the above table that the minimum emission of the benzoghiperylene chemical is 0 and the maximum emission is 65112.8. Its average emission is 15.04 with a standard deviation of 515.28. We would like to perform a more in-depth exploratory analysis and present our findings in the next assignment.

Subgroups

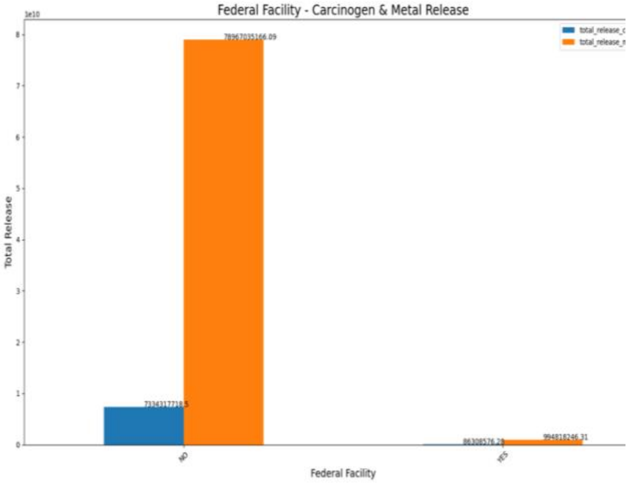
It would be more meaningful to cluster data and perform our tests to make better predictions. We would like to group the tri facility data by the following attributes and perform exploratory, inferential, and predictive analysis.

- Year
- Company Name
- Federal Facility
- Parent company Name
- Country
- State
- City

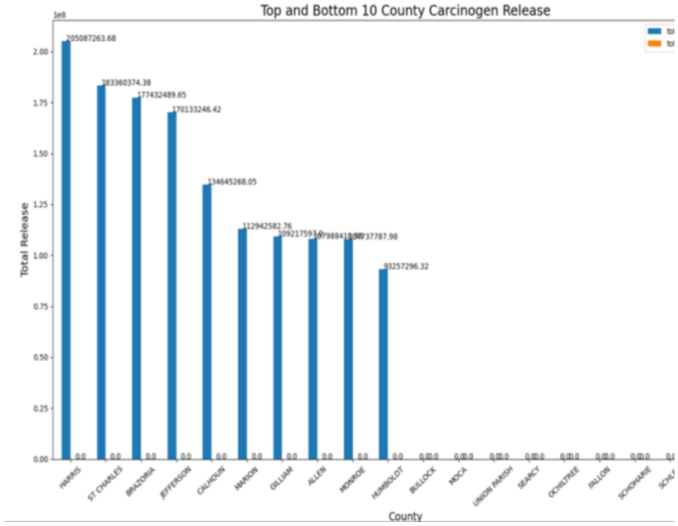
3) Exploratory Analysis:



From the above Bar plot, you can see the total release of carcinogen and metal in the years 1987-2014. While we see that there is a significant drop in the levels of carcinogens released in 1989 as compared to 1988, the low levels continued up until 1997, after which rapid industrialization began, leading to higher emission of carcinogens and release of metals. The year 2014 has the highest number of Metal releases compared to all the years whereas 2012 has the lowest number of carcinogens and metal releases.



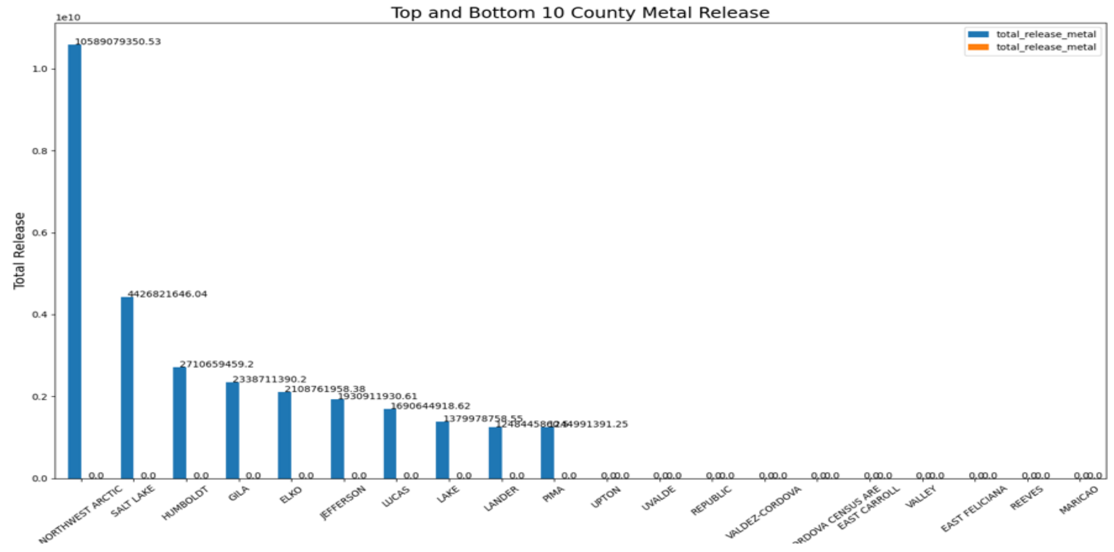
From the plot, you can say that the places which have federal facilities have lower amounts of carcinogens and metals released. The places which do not have federal facilities have the highest number of carcinogens and metals released when compared to the places which have federal facilities. The total amount of metal released is very much higher than the total amount of Carcinogens released in places that do not have federal facilities. From the above plot you can conclude that the places having federal facilities have less carcinogens and metal releases.



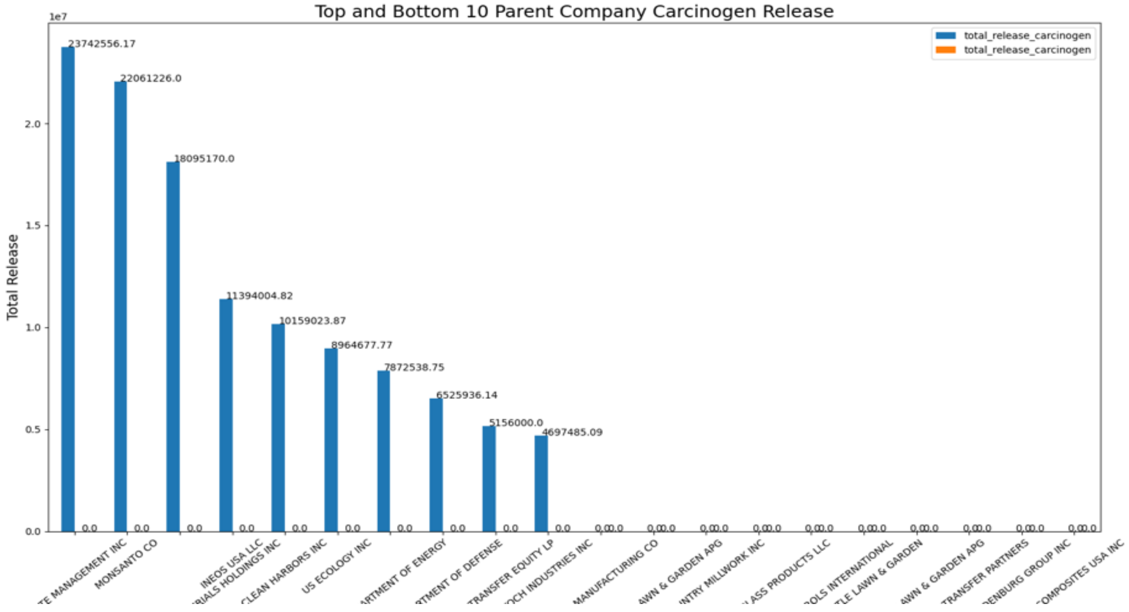
ss

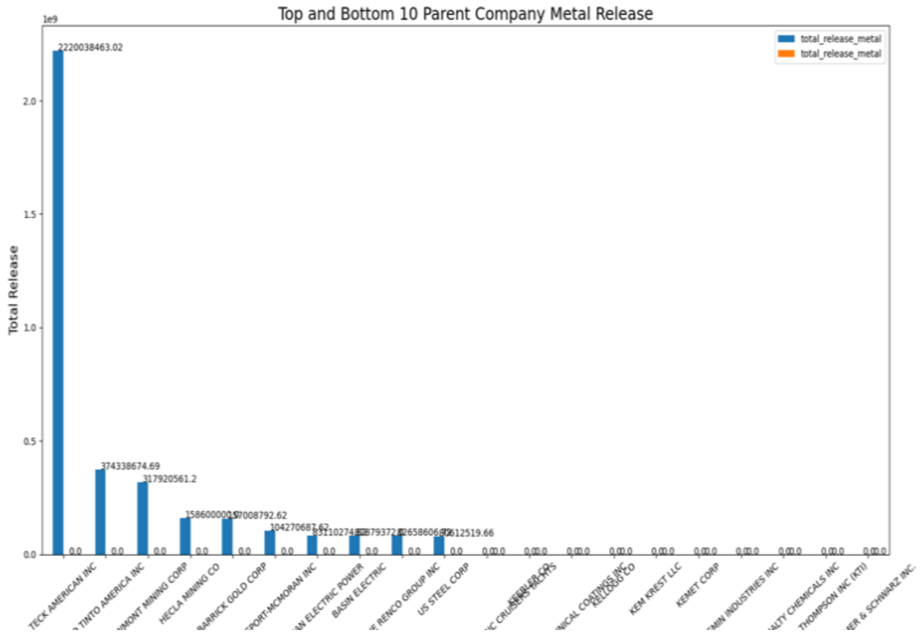
From the plot, you can see the top and bottom 10 counties with total amount of carcinogens released by each county. Harris County has the highest number of carcinogens released followed by St. Charles. The Humboldt County has the lowest number of carcinogens released compared to the other counties. And from the plot you can see that all the bottom 10 counties have zero release of carcinogen in all the years.

From the below plot, you can see the top and bottom 10 counties with total amount of metals released by each county. North-west Artic has the highest number of Metals released followed by Salt Lake. Pima County has the lowest number of Metals released compared to the other counties. And from the plot you can see that all the bottom 10 counties like Upton, Republic, Valley etc. have zero release of metals in all the years.



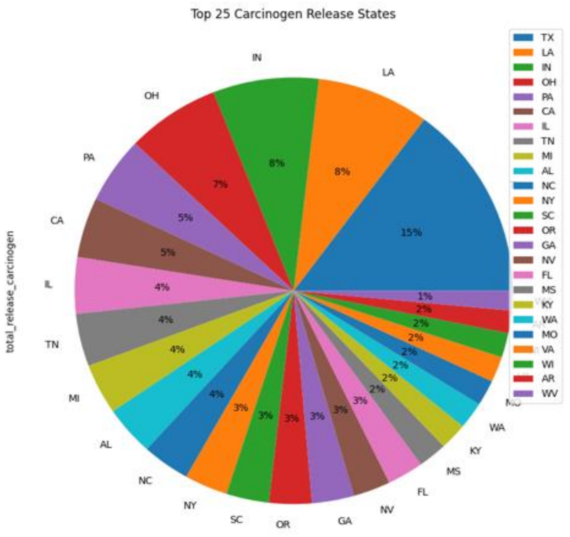
From the plot below, you can see the top and bottom 10 Parent companies with total amount of Carcinogen released by each parent company. Waste Management Inc has the highest number of Carcinogens released followed by Monsanto co. Transfer equity Lp has the lowest number of carcinogens released compared to the other companies. And from the plot you can see that all the bottom 10 parent companies like Manufacturing co, Composites USA Inc, etc. have zero total release of carcinogens in all the years.

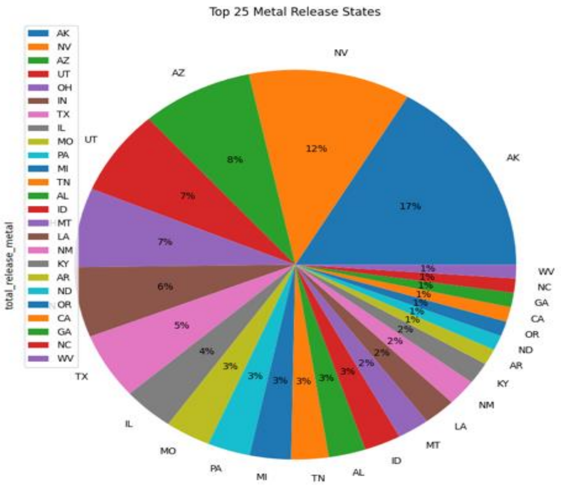




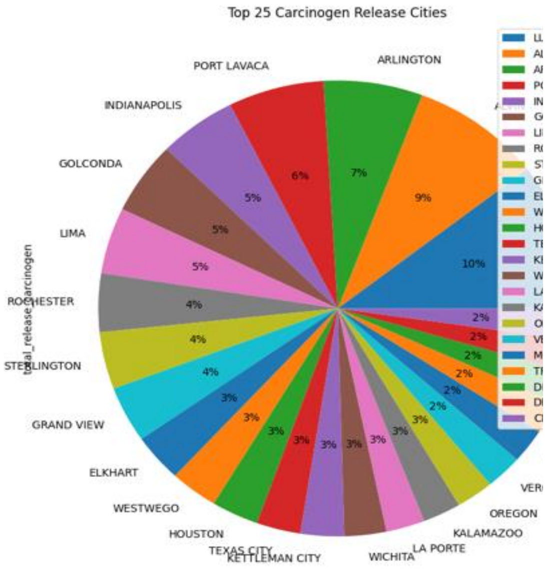
From the plot, you can see the top and bottom 10 Parent companies with total amount of Metals released by each parent company. Tech American Inc has the highest number of Metals released followed by Rio America Inc. The Renco Group has the lowest number of Metals released compared to the other parent companies. And from the plot you can see that all the bottom 10 parent companies like Kemet Corp, Kem Krest LLC etc. have zero total release of Metals in all the years.

From the pie chart, you can see the top 25 Carcinogen release states. Texas has the highest percentage of Carcinogen release which is 15% followed by Louisiana which is 8% followed by Indiana which is 8% as well. West Virginia has the lowest percentage of Carcinogen release which is 1% followed by Arkansas which is 2% followed by Wisconsin which has 2% which Is very negligible compared to all the other states.

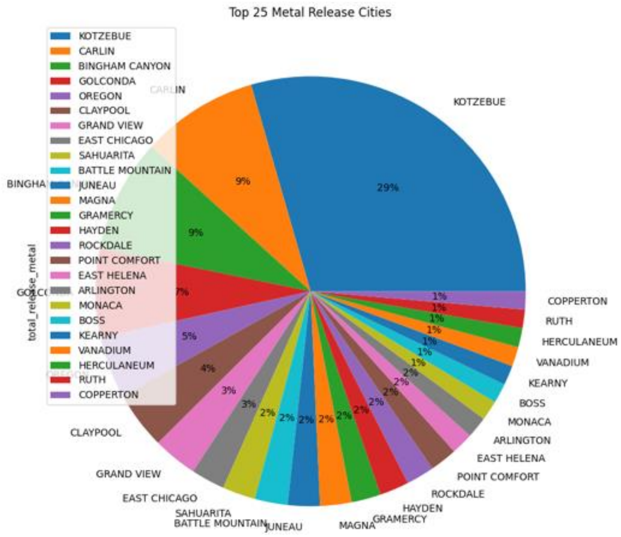




From the pie chart, you can see the top 25 Metal release states. Arkansas has the highest percentage of Metal release which is 17% followed by Nevada which is 12% followed by Arizona which is 8%. West Virginia has the lowest percentage of Metal release which is 1% followed by North Carolina which is 1% followed by Georgia which has 1% which Is very negligible compared to all the other states. From the above two pie charts you can see that West Virginia has the lowest percentage of carcinogens and metals released when compared to other states.

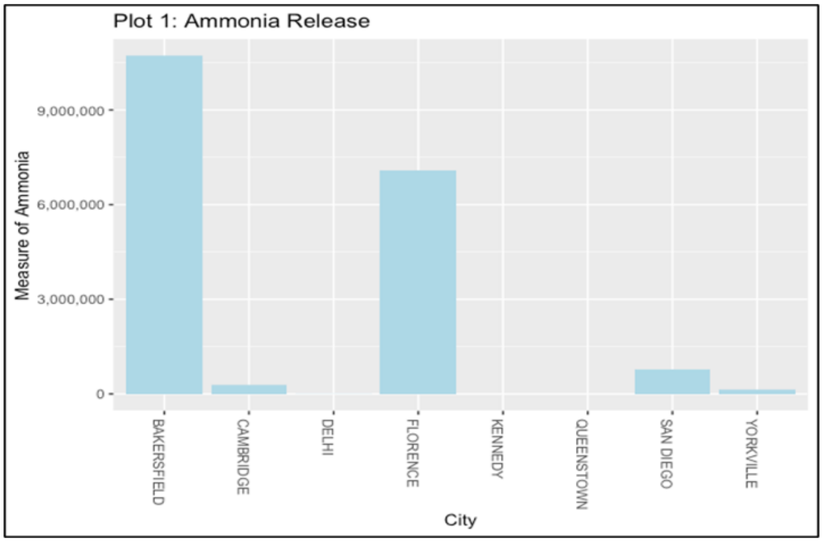


From the above pie chart, you can see the top 25 Carcinogen release Cities. Luling has the highest percentage of Carcinogen release which is 10% followed by Alvin which is 9% followed by Arlington which is 7%. Clinton, Decatur, Deer Park, Troy, Mount Vernon have the lowest percentage of Carcinogen release which is 2% followed by Oregon, Texas, etc. which are 3% which Is negligible compared to all the other Cities

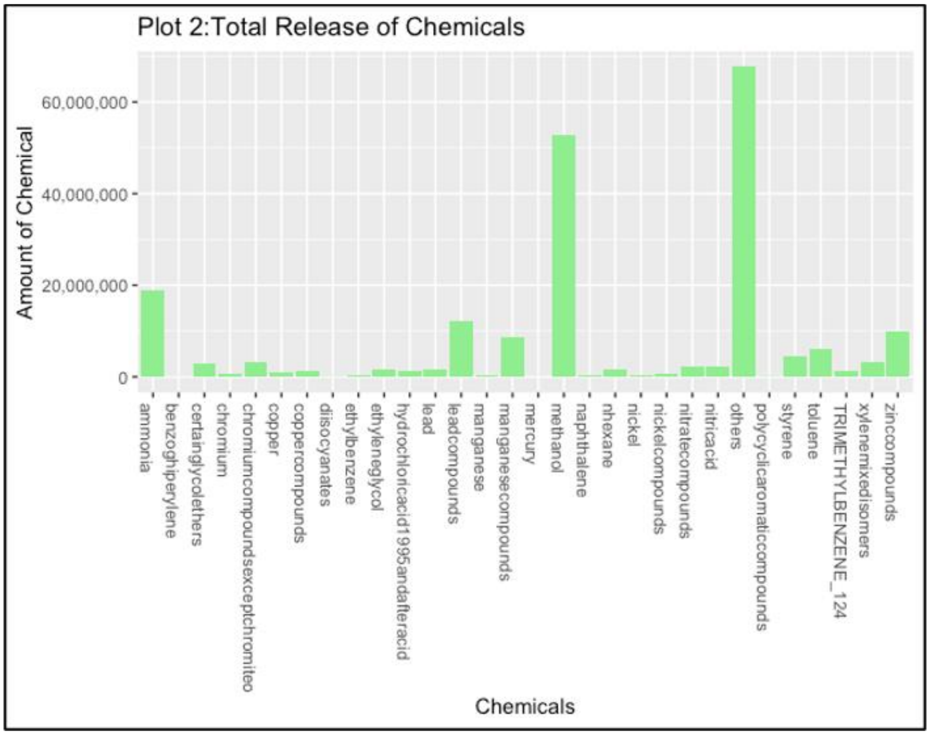


From the above pie chart, you can see the top 25 Metal release Cities. Kotzebue has the highest percentage of Metal release which is 29% followed by Carlin which is 9% followed by Bingham Canyon which is 9%. Copperton, Ruth, Vanadium, Kearny, Boss, Monaca have the lowest percentage of Metal release which is 1% followed by Arlington, Hayden etc. which are 2% which Is negligible compared to all the other Cities.

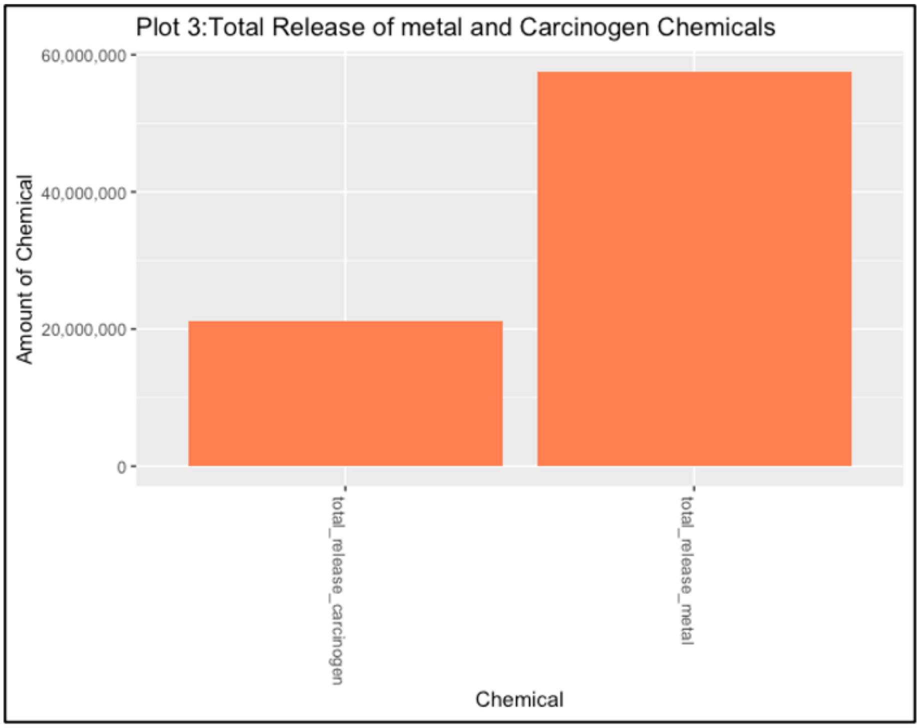
Let us look at the plot below which gives us the visual representation of the total ammonia released in each of the eight cities. As we can see, Bakersfield is the highest and Queensland and Kennedy barely release any ammonia. Florence is the city which releases the second highest amount of ammonia.

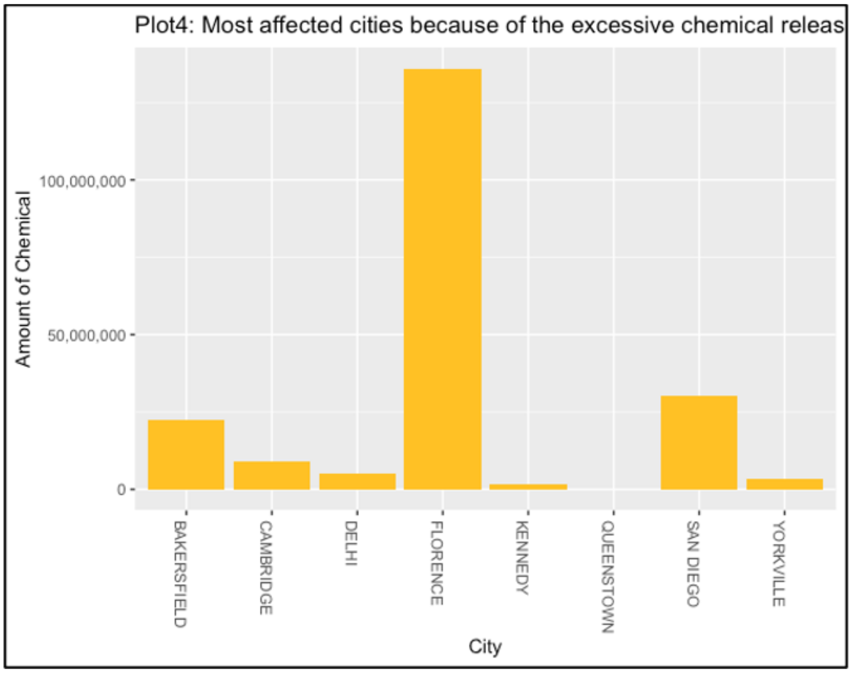


The below graph gives us an estimate of all the chemicals released in the eight cities put together. As we can see, the sum of the rest of the chemicals is the highest followed by methanol and ammonia and then lead compounds. Some of the least released chemicals are benzoghiperylene, diisocyanatos, mercury, and polycyclic aromatic compounds.



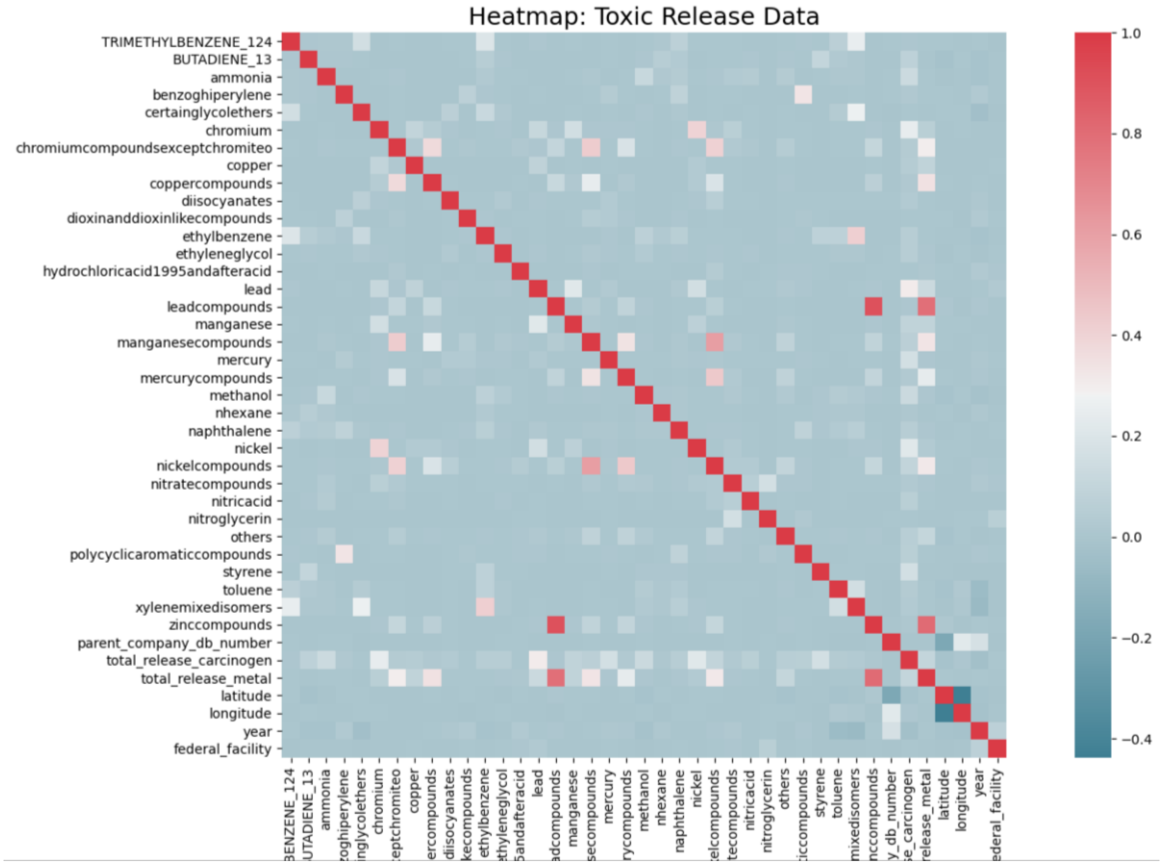
Above is the plot showing the total amount of carcinogen and metal released. The total metal released is approximately 60,000,000, which more that the double of total carcinogen released which is roughly 20,000,000.





The above plot shows us the most affected cities due the total chemical release. As we can see Florence is the most affected city which is more than five times the following highest affected city, Bakersfield. Queensland is the least affected city followed by Kennedy.

4) Correlation Analysis



From the above plot, it is evident that dioxinanddioxinlikecompounds and manganesecompounds are strongly positively correlated with the attribute total_release_metal. Also, leadcompounds is positively correlated with manganesecompounds and total_release_metal. We are interested in fitting a model for the total_release_metal and so either of the manganesecompounds and leadcompunds can be used as a primary supporting independent attribute. We are interested to fit a model for total_release_carcinogen attribute but identified any attribute with positive correlation with it. We could identify the top supporting attributes with the help of stepwise regression.

5) Hypothesis Testing

Here, we want to check that Carcinogen and Metal release is same for all when we group results by parental company, year, federal facility, state, and city. Let us evaluate the hypothesis test results below.

T and Z tests

Null Hypotheses: Both Carcinogen and Metal release same amount of toxic chemicals.

Alternative Hypotheses: Both Carcinogen and Metal release different amounts of toxic chemicals.

Significance Value: 0.05

Grouped by parental company:

Statistics	T-test	Z-test
Test score	-2.5691	-2.5691
P value	0.0102	0.0101
Status	Rejected	Rejected

Grouped by county:

Statistics	T-test	Z-test
Test score	-5.4689	-5.4689
P value	0.0000	0.0000
Status	Rejected	Rejected

Grouped by state:

Statistics	T-test	Z-test
Test score	-4.4565	-4.4565
P value	0.000041	0.000007
Status	Rejected	Rejected

Grouped by city:

Statistics	T-test	Z-test
Test score	-5.7631	-5.7631
P value	0.0000	0.0000
Status	Rejected	Rejected

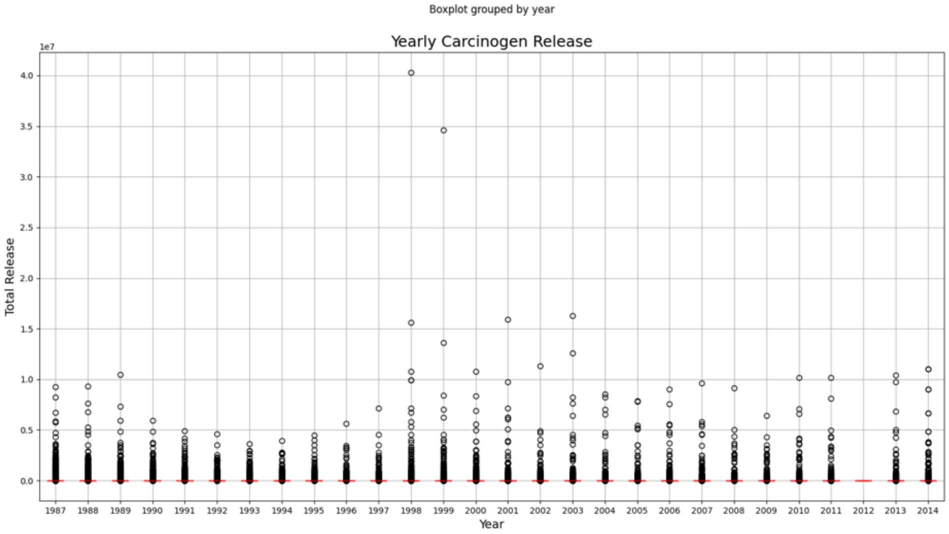
After performing t and z tests, we can see that the p-values and z-values are lesser than the significant value which is 0.05. At 95% confidence level, we can reject the null hypothesis which states that the Carcinogen and Metal release same amount of toxic chemicals. The levels of these chemicals released are different, which may be due to the location of the facilities, or the type of work carried out by the facility or the federal facilities of the states.

ANOVA:

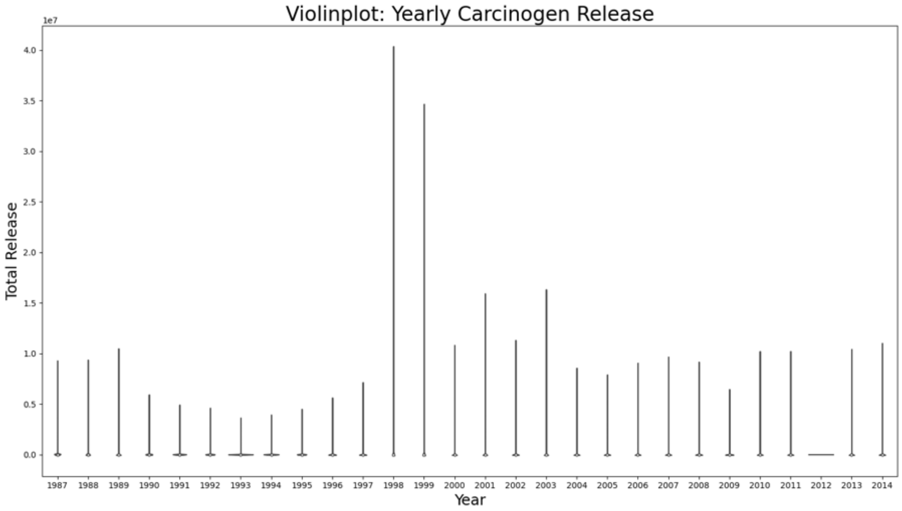
One Way ANOVA

total_release_carcinogen ~ year

Bos and Violin plots

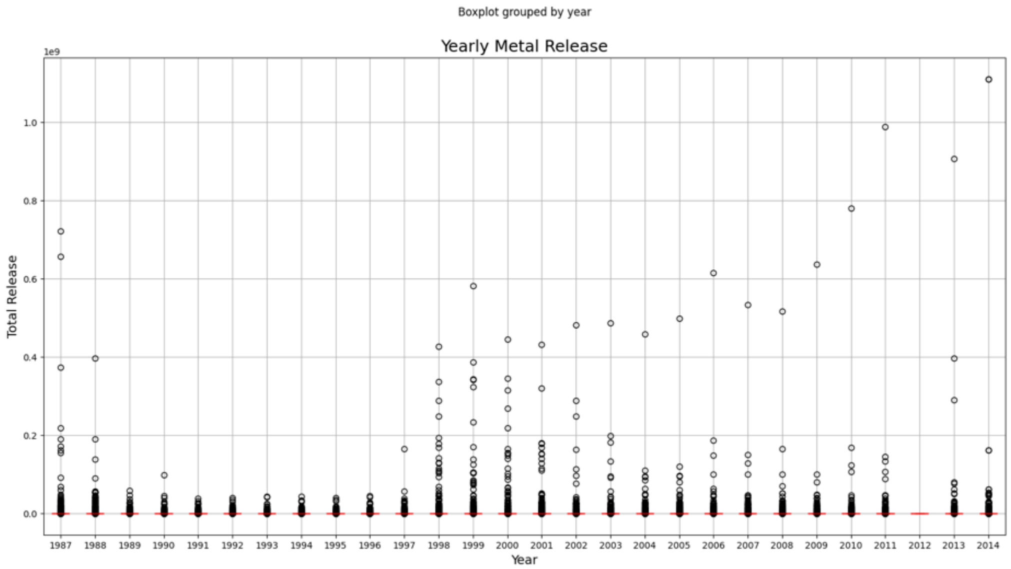


We can see in the above plots that the total Carcinogen release is consistent in the earlier years and has extreme outliers in the years 1998 and 1999. It has a few significant outliers in the years 2001, 2002 and 2003 also.

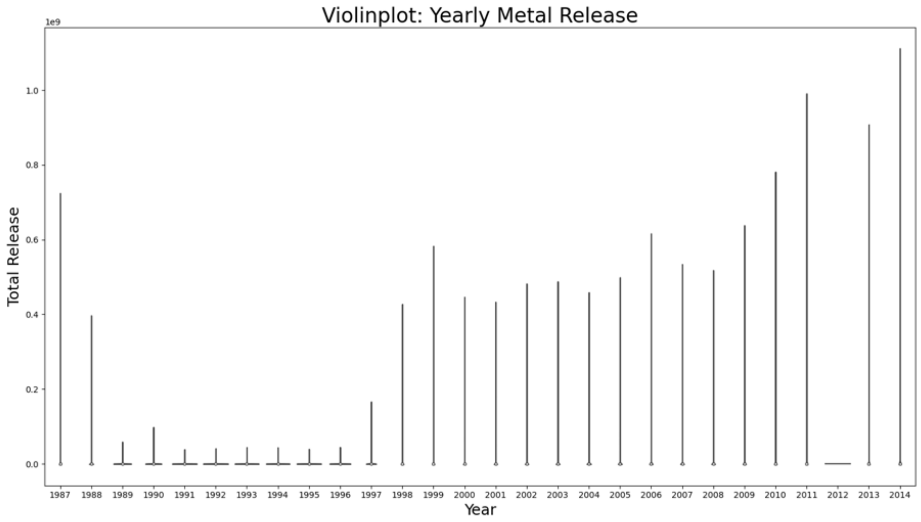


total_release_metal ~ year

Bos and Violin plots



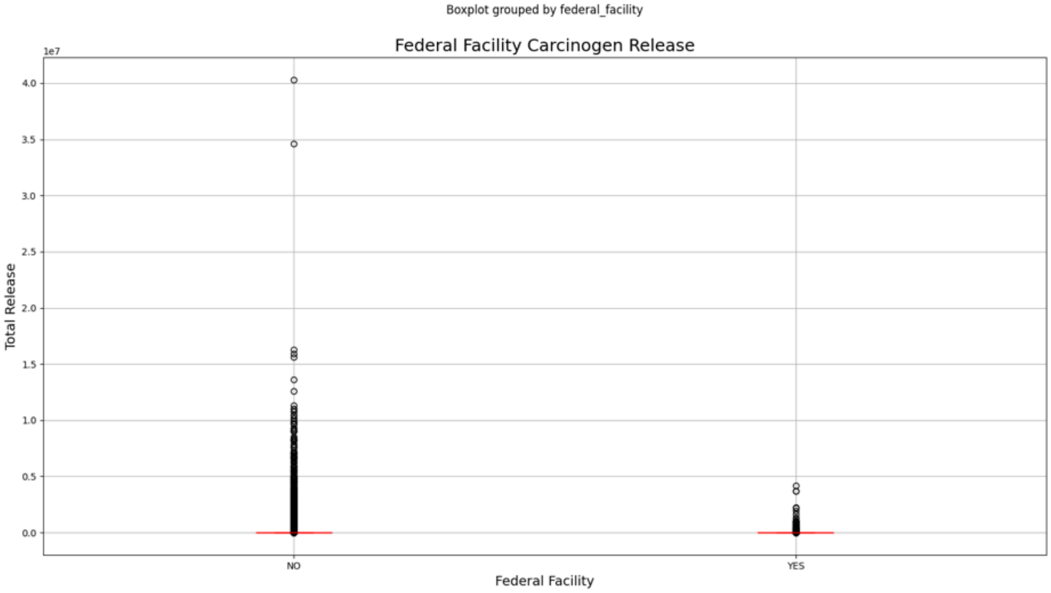
Let us look at the boxplot grouped by year for the total metal release. The first two years, 1987 and 1988, have a few outliers and then the distribution is consistent. We can again see quite a few significant outliers starting from the years 1998 to 2014. Out of all the years, 2011 and 2014 have extreme outliers.



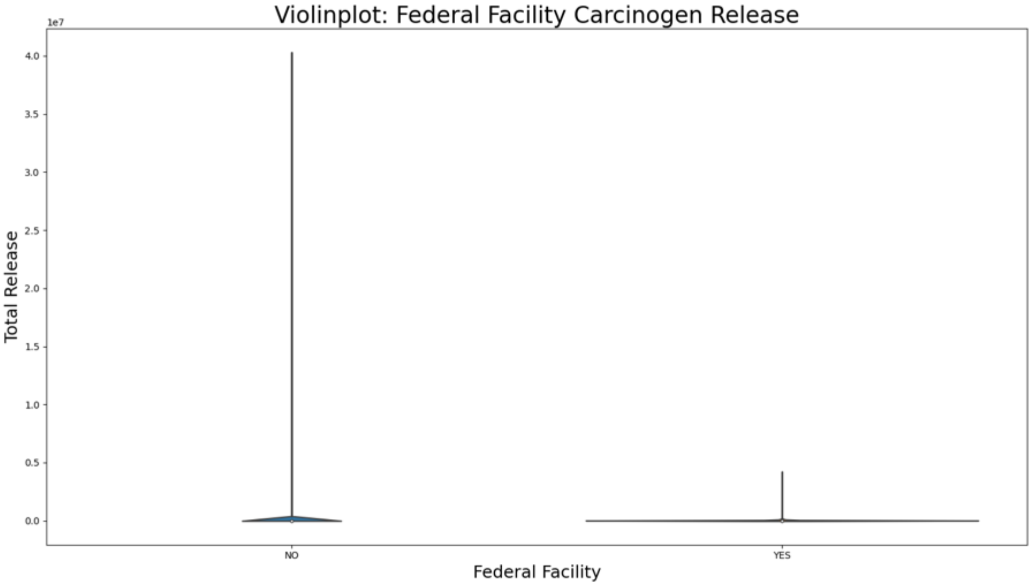
Statistics	total_release_carcinogen	total_release_metal
sum_sq	9.91*e+12	2.76*e+14
df	1.0	1.0
F	512.22	14.27
PR(>F)	2.29*e-113	0.000158

total_release_carcinogen ~ federal_facility

Bos and Violin plots

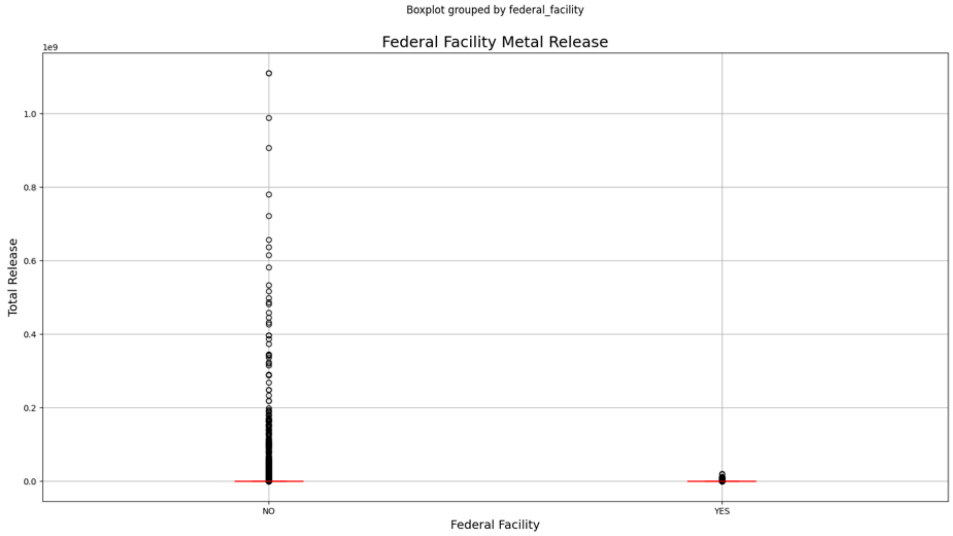


The above boxplot is of the total carcinogen released, grouped by the federal facility. The first one has outliers while the second one seems to be distributed normally.

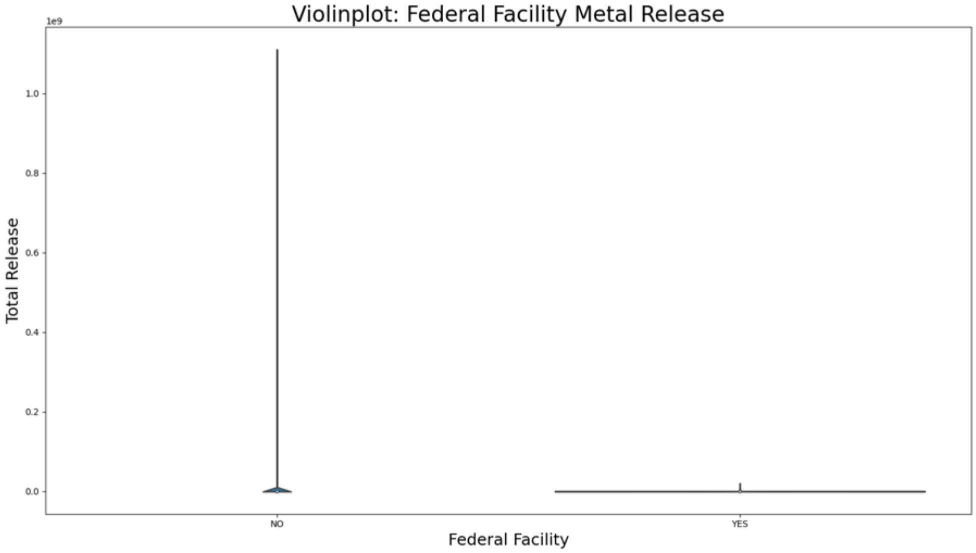


total_release_metal ~ federal_facility

Bos and Violin plots



Let us look at the boxplot of the total metal release grouped by the federal facility. The first one has outliers while the second one seems to be distributed normally.



Statistics	total_release_carcinogen	total_release_metal
sum_sq	2.27*e+10	5.82*e+12
df	1.0	1.0
F	1.17036	0.301091
PR(>F)	0.279328	0.5832

Two Way ANOVA:

total_release_carcinogen ~ C(year) + C(federal_facility) + C(year):C(federal_facility)

Attribute	sum_sq	df	F	PR(>F)
C(year)	1.45*e+13	27.0	27.7466	0.0000
C(federal_facility)	1.12*e+11	1.0	7.3409	0.0067
C(year):C(federal_facility)	1.54*e+11	27.0	0.2952	0.9999

a. H0: There is no difference in total release of carcinogen with respect to year

   H1: There is a difference in total release of carcinogen with respect to year

b. H0: There is no difference in total release of carcinogen with respect to federal facility

   H1: There is a difference in total release of carcinogen with respect to federal facility

c. H0: There is no interaction between year and federal facility

   H1: There is interaction between year and federal facility

In the first two cases the p-values are less than the significant values, therefore we can reject the null hypothesis mentioned above. In the last case the p-value is 0.9999 which is greater than the significant value, therefore we do not have enough evidence to reject the null hypothesis.

total_release_metal ~ C(year) + C(federal_facility) + C(year):C(federal_facility)

Attribute	sum_sq	df	F	PR(>F)
C(year)	3.41*e+15	27.0	6.5409	0.0000
C(federal_facility)	2.02*e+12	1.0	0.1044	0.7467
C(year):C(federal_facility)	1.76*e+13	27.0	0.0338	1.0000

a. H0: There is no difference in total release of metal with respect to year

   H1: There is a difference in total release of metal with respect to year

b. H0: There is no difference in total release of metal with respect to federal facility

   H1: There is a difference in total release of metal with respect to federal facility

c. H0: There is no interaction between year and federal facility

   H1: There is interaction between year and federal facility

In the first case the p-value is less than the significant value, therefore we can reject the null hypothesis mentioned above. In the last two cases the p-values are greater than the significant value, therefore, we do not have enough evidence to reject the null hypothesis.

6) Data Scaling & Encoding

Given data has uniform values, so imbalance is observed and so scaling is not required.

Federal Facility is encoded for training purposes as it has good correlation.

7) Data Splitting

Given data is split into two parts. 70% is for training and 30% is for testing purposes.

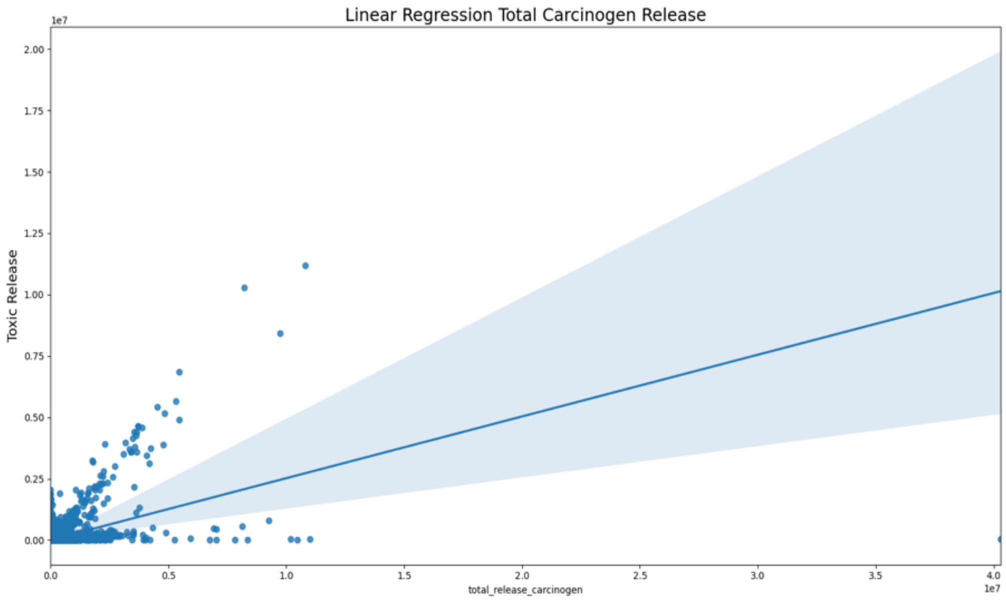
8) Regression Analysis

Here, we would like to fit a total of three regression models:

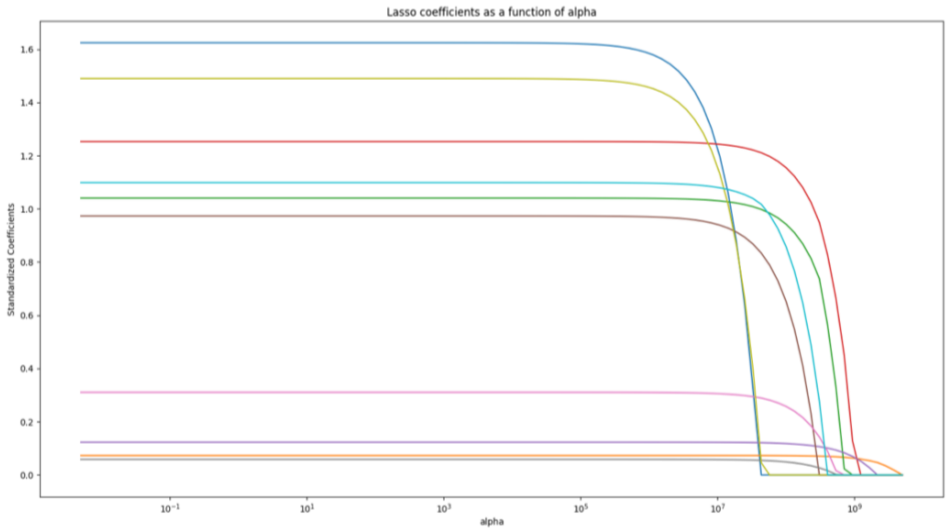
1) Linear Regression model for the prediction of carcinogen toxic release.

We have used the forward step wise selection technique to identify the most correlated supporting variables and fitted linear regression models using them. Below are the identified supporting variables and the regression plot.

['BUTADIENE_13', 'ammonia', 'chromium', 'lead', 'methanol', 'nickel', 'nickelcompounds', 'nitricacid', 'polycyclicaromaticcompounds', 'styrene']

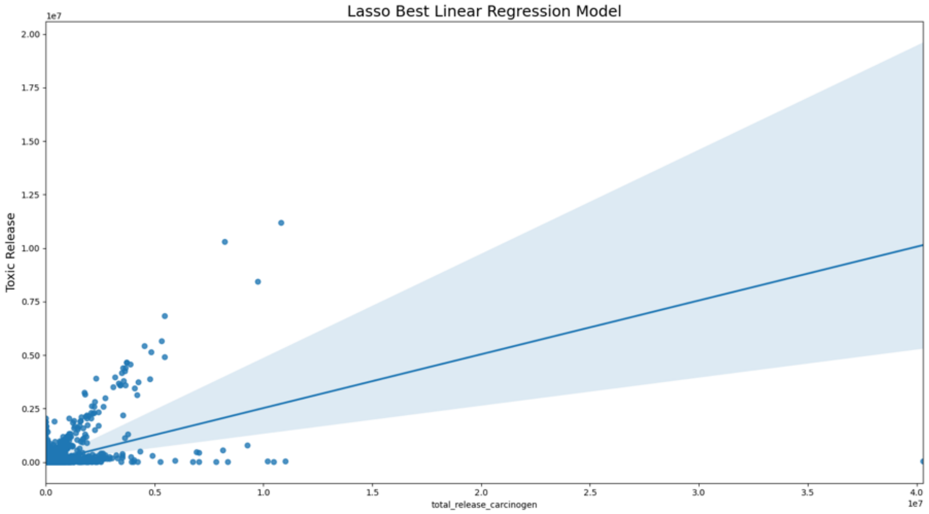


The next step is to use regularization techniques to handle overfitting. The above plot shows the model is under-fitted but let us look at the lasso regression fitting below. First, let us find out the best alpha parameter.



The best alpha is identified as 0.99 after running for 100 iterations.

Let us check the lasso regression model below.



The plot looks like the previous one. Let us check the RMSE values and decide on their improvement.

Step Wise Regression RMSE: 127878.45799329084

Lasso Regression RMSE: 127878.4579910477

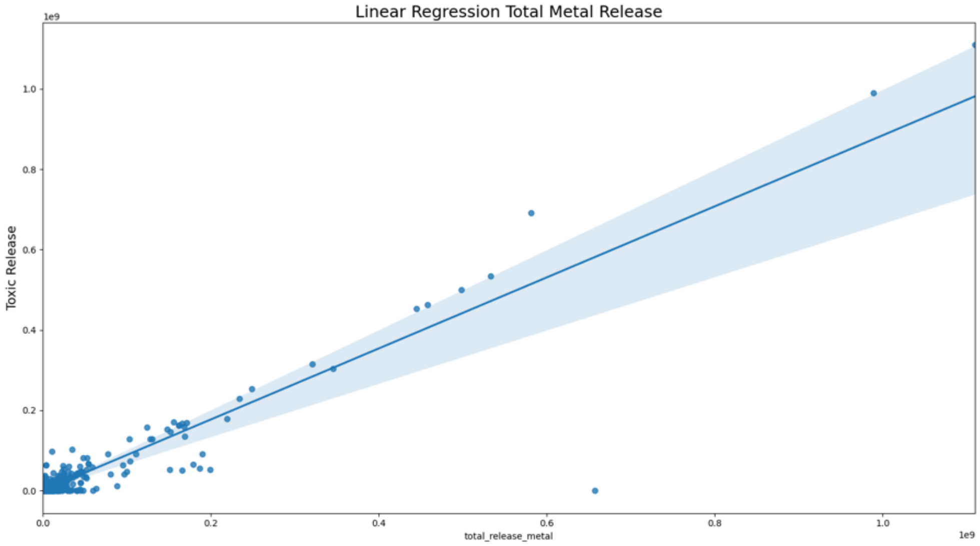
We could see an improvement at the 6th decimal value which is not helpful. This is expected because data does not have any variable with a correlation value of greater than 0.5. In fact, all the top ten correlated variables correlation values are around 0 which means neutral.

Let us go ahead and repeat the same process for the total release metal target variable.

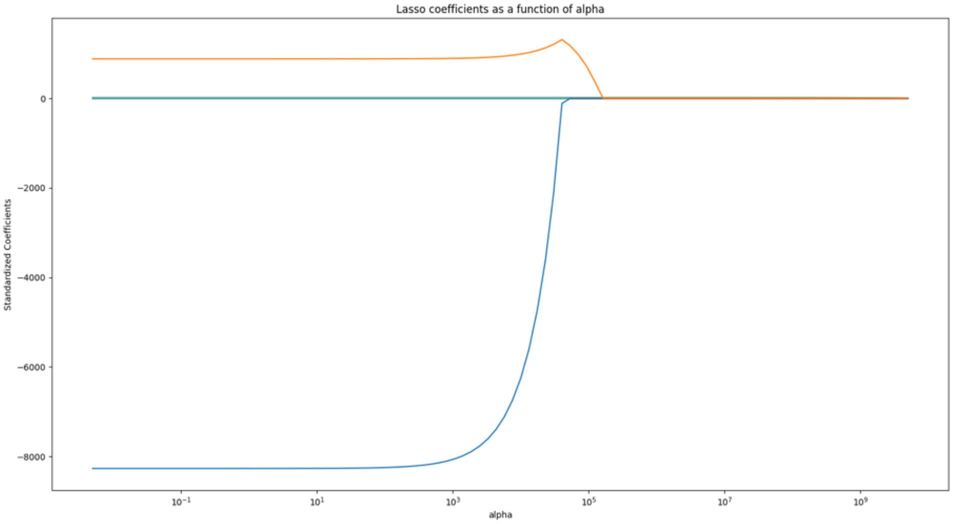
2) Linear Regression model for the release of metal toxic release.

We have used the forward step wise selection technique to identify the most correlated supporting variables and fitted linear regression models using them. Below are the identified supporting variables and the regression plot.

['chromium', 'chromiumcompoundsexceptchromiteo', 'copper', 'coppercompounds', 'lead', 'leadcompounds', 'manganesecompounds', 'mercurycompounds', 'nickelcompounds', 'zinccompounds', 'latitude', 'longitude']

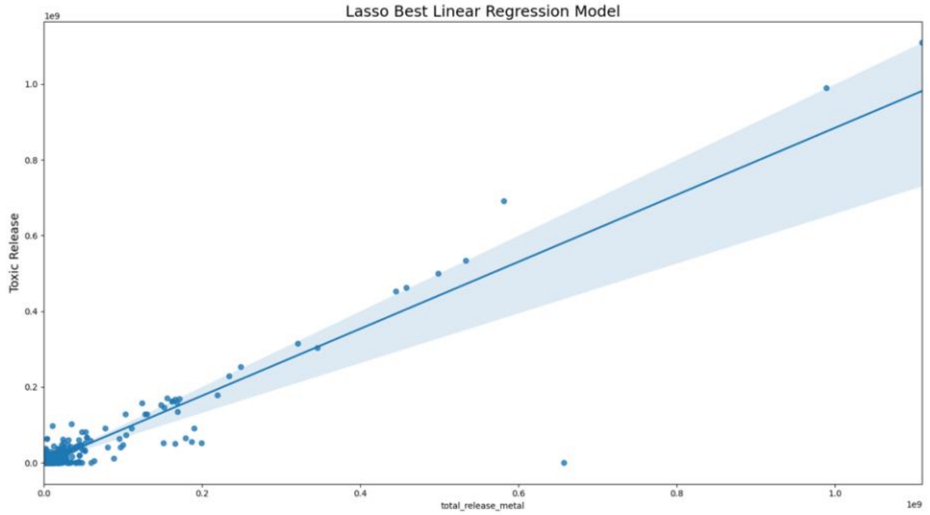


The next step is to use regularization techniques to handle overfitting. The above plot shows the model is under-fitted but let us look at the lasso regression fitting below. First, let us find out the best alpha parameter.



The best alpha is identified as 0.98 after running for 100 iterations

Let us check the lasso regression model below.



The plot looks like the previous one. Let us check the RMSE values and decide on their improvement.

Step Wise Regression RMSE: 1791410.8059675016

Lasso Regression RMSE: 1791410.8066408692

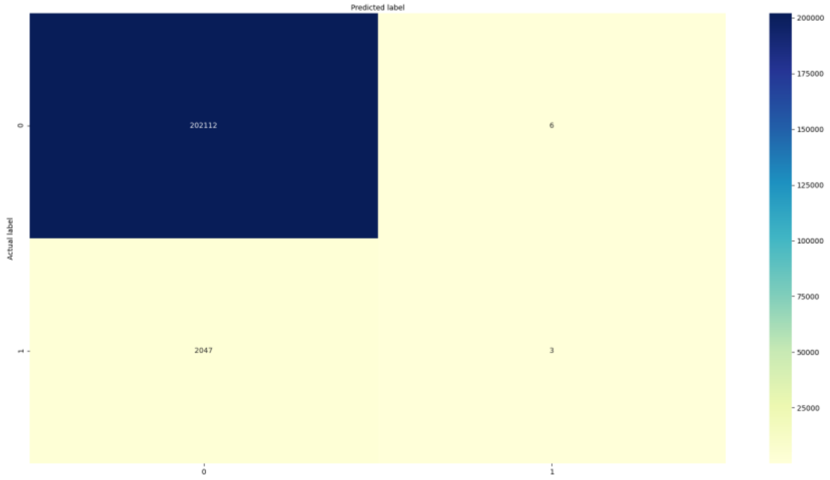
We could see an improvement at the 6th decimal value which is not helpful. This is expected because data does not have any variable with a correlation value of greater than 0.5. In fact, all the top ten correlated variables correlation values are around 0 which means neutral.

3) Logistic Regression model for the prediction of federal facility availability.

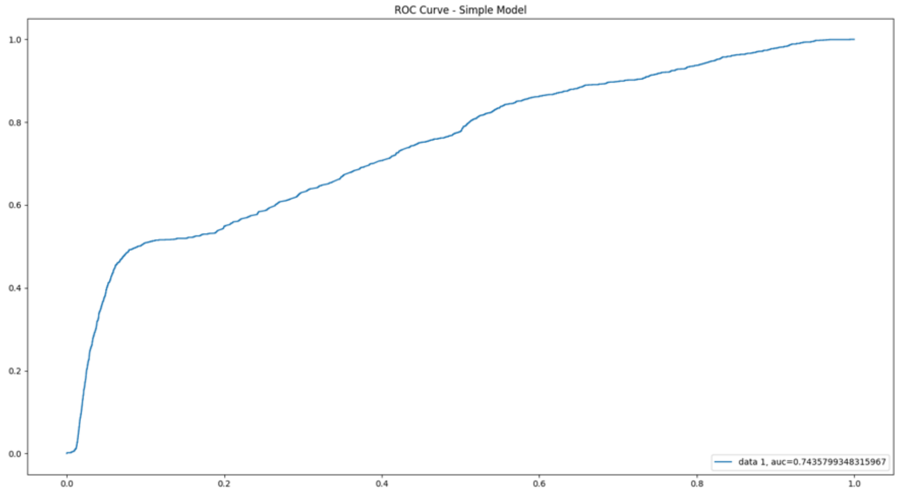
We have fitted the logistic regression model and observed the precision and recall values are bad. We suspect that this has happened because of the imbalance in the dataset.

Accuracy: 98.99%

Precision: 33%



Though accuracy is 98 percent this model cannot be used in production as its precision and recall values are bad.



The above is the total area under the curve. The higher the area better is the performance of the model. It is evident from the plot that the top left portion is empty which is another sign that the dataset has imbalance classes in the federal_facility attribute.

Let us go ahead and build high level models like LGB and XGB to see if those models can internally handle this data imbalance and fit well.

9) High Level Models Building

We have successfully trained Light Gradient Boosting, Extreme Gradient Boosting, Random Forests, SVM, and Naive Bayes models and tabulated their evaluation metrics below. Every model was trained on the 70% data and the rest 30% was used for testing purposes.

Model	Accuracy	Precision	Recall	F1 Score
Light Gradient Boosting	**98.24**	**98.14**	**98.11**	**98.12**
Extreme Gradient Boosting	**97.56**	**98.12**	**96.88**	**97.64**
Random forest	**99.24**	**99.18**	**99.37**	**99.29**
SVM	**99.99**	**99.99**	**99.99**	**99.99**
Naive Bayes	**98.99**	**99.00**	**98.00**	**99.00**
Logistic Regression	**98.99**	**33.33**	**0.24**	**0.36**

10) Models Evaluation

As expected, all-time great tree-based boosting models handled this class imbalance internally and fitted well on the given data. We could see the consistency in accuracy, precision, recall, and F1 score metrics. All values are around 98% percent which means fitted well. We could tune regularization parameters to penalize the overfitting. Going down to the table, we could see that metric values are greater than 99% for random forests, svm, and naïve bayes which means these models could not handle the class imbalance and were overfitted. The Logistic regression model is the only one that was underfitted on the given data. We have tried using step wise regression and lasso regression for logistic regression but none of them helped to fit well on the data.

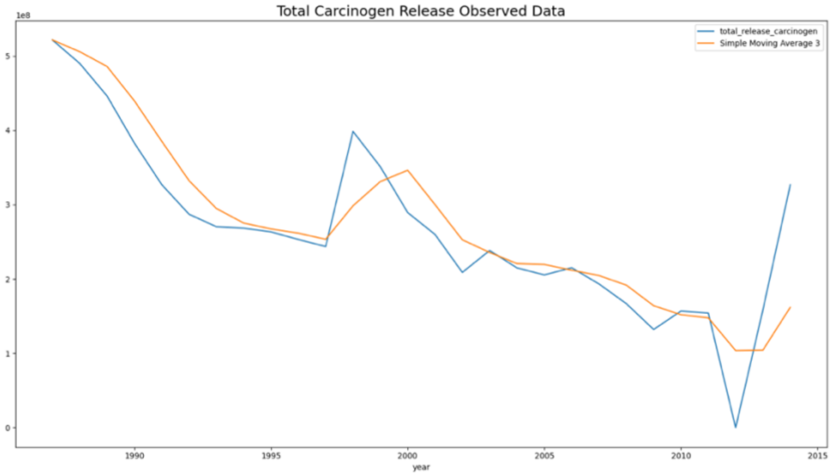
We have personally experienced why light gradient boosting and extreme gradient boosting models are called champion models especially for famous competitions like Kaggle and Hacker earth. We are good at using the lgb and xgb models for predictions.

11) Predictive Analytics

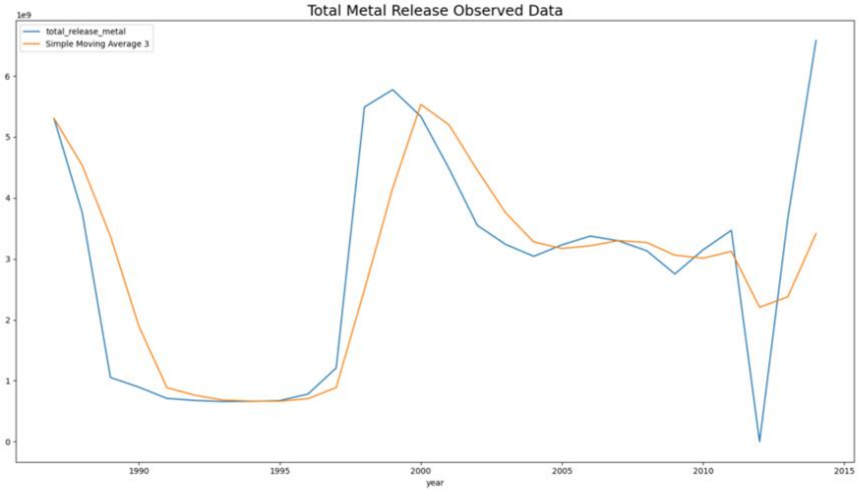
We have used companies' facilities dataset for predictions and seen both lgb and xgb giving realistic prediction values.

12) Time Series Analysis

Since the given dataset has a time feature which year, we thought of fitting time series model and check the components like trend, seasonality, and randomness for both carcinogen metal total yearly releases.

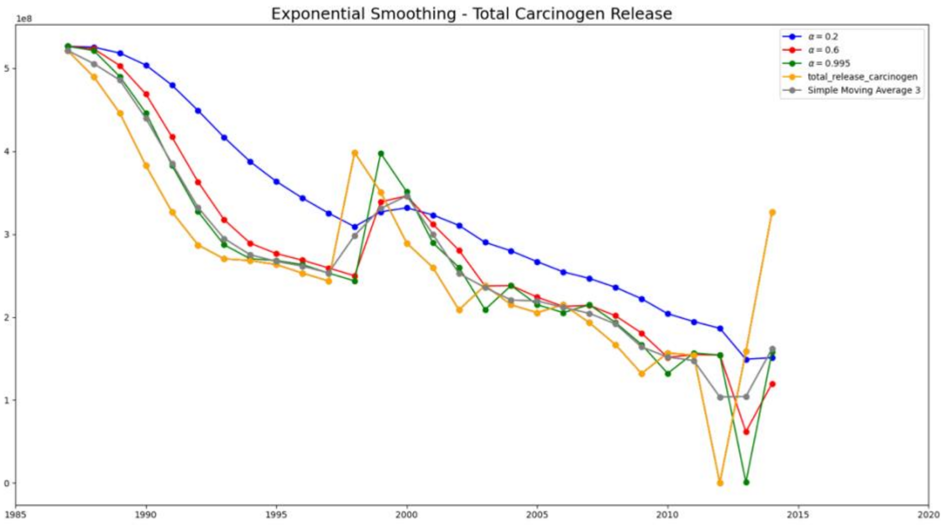


The above is the yearly carcinogen total release data and its three-period moving averages. We could see that there has been a decrement in toxic releases over the years. We can also observe that trend direction started changing which gives us a sign that toxic release would be increased which can be forecasted using exponential smoothing model below.

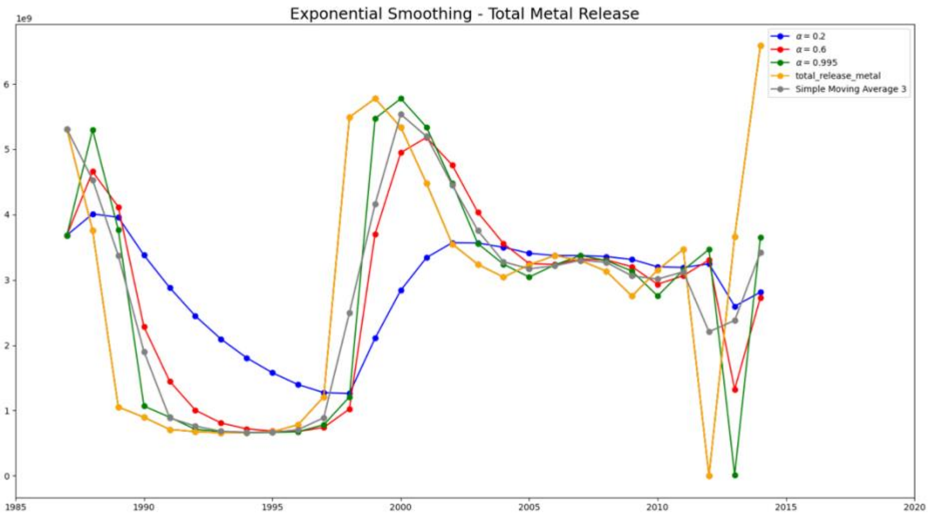


The above is the yearly metal total release data and its three-period moving averages. We could see that there was a sudden drop in toxic releases in 1990 and it continued the same till 1998. After that, there was a sudden increment and then started decreasing till 2014. Like the Carcinogen plot, the above-trend direction started changing which gives us a sign that toxic release would be increased which can be forecasted using the exponential smoothing model below.

Let us go ahead, build exponential smoothing models, and make forecasts for the next 5 years.



The above plot is the forecasted values of carcinogen release from the year 2014 to 2019. We could see predictions at three different smoothing levels – 0.2, 0.6, and 0.995. We could see model fitted with smoothing parameter of 0.995 represented the data well, so we can use its forecasted values. However, this model tells us that there would be a sudden drop after 2014 and then a sudden increment which can be confirmed with real data.



The above plot is the forecasted values of metal release from the year 2014 to 2019. We could see predictions at three different smoothing levels – 0.2, 0.6, and 0.995. We could see model fitted with smoothing parameter of 0.6 and 0.995 represented the data well, so we can use either of its forecasted values. However, this model tells us that there would be a sudden drop after 2014 and then a sudden increment which can be confirmed with real data.

1) Find the highest amount of chemicals released each year, and predict its trend?
A) The sum of the rest of the chemicals is the highest followed by methanol and ammonia and then lead compounds.

2) Which facility releases the highest amount of toxic compounds in total each year?
A) Waste Management Inc has the highest number of Carcinogens released followed by Monsanto co.

3) Which compounds have the highest correlation?
A) Dioxinanddioxin like compounds and manganesecompounds are strongly positively correlated with the attribute total_release_metal.

4) Which compounds have the lowest correlation?
A) Total carcinogen release has the lowest correlation with lead and chromium. Total metal release has the lowest correlation with mercury compounds.

5) Find the Total Release of top 30 Chemicals in all Years?
A) Methanol has been released in quantities up to 52,000,000, making it the most widely released chemical, followed by ammonia, which has been released in quantities up to 20,000,000.

6) Find the Total Release of metal and Carcinogen Chemicals?
A) Total amount of carcinogen released over all the years is approximately 20,000,000 and that of metal is 60,000,000

7) Find the Highest number of Chemicals Released by each State?
A) Texas has the highest number of carcinogen chemicals emitted, accounting for 25% of the total in the US, while Arkansas has the highest number of metal releases, accounting for 17%.

8) Find the Lowest number of Chemicals Released by each State?
A) In the United States, West Virginia has the lowest quantity of carcinogen chemicals and metals discharged, with less than 1%.

9) Does every county have the same threat?
A) No, county with federal facilities have low threat compared to others.

10) What are the most affected cities because of the excessive chemical releases?
A) Most affected city due to Carcinogen release is Luling, followed by Alvin. Most affected city due to Metal release is Kotzebue followed by Carlin and Bingham Canyon

11) Is there any seasonal pattern behavior associated with any company?
A) Yes, chemical releases have seasonality and downtrend associated with them.

12) What are the causes of excessive chemical releases and how to prevent them?
A) Not using quality raw products is one of the causes for the excessive chemical releases.

13) How to build the best pipeline using analytical methods for identifying the threats and prevent them early?
A) Any pipeline of analytical approach with support from numerical approach provides a reliable solution. The pipeline that we have used worked well for identifying the threats in advance.

Conclusion:

- Given dataset has missing values which we imputed with mode and mean values. We have successfully completed the exploratory analysis and hypothesis testing on the given data and observed that the total amount of carcinogen and metal release is different in the subgroups like County, Parental Company, State, City, and Year. The same has been mathematically supported with T and Z tests and ANOVA tests. We have successfully identified the top affected cities and states in each year and generated the visual plots of the same.
- We have built two linear independent regression models for predicting the total release of Carcinogen and Metal. The model for the metal release prediction fitted well compared to carcinogen release prediction. This is because the metal release feature has a minimum of two strongly positively correlated supporting features whereas the carcinogen release feature has negatively correlated features. We have tried to build a logistic regression model for the federal facility prediction, but it did not fit well because of the class imbalance. Hence, we went on building all the advanced machine learning traditional models like LGB, XGBoost, Random Forests, SVM, and Naïve Bayes. We have observed that only LGB and XGBoost models were able to handle the class imbalance and fitted well with evaluation metric scores of greater than 98 percent.
- We have observed a trend in the total amount of toxic chemical releases over the years from 1987 to 2014, so wanted to fit time series models. We have grouped the data by year, fitted the Simple Exponential Smoothing model on both total release carcinogen and metal features, and made forecasting for a period of 5 years. Carcinogen total release has a uniform downtrend from 1987 to 2014 whereas metal total release has a sudden drop in the 1990 year and then a sudden spike in the 2000 year, and from there trend started falling uniformly. In both cases, modes fitted at a smoothing factor of 0.995 best represented the data and provided more realistic forecasting values.
- **Future Scope:** We want to use advanced sampling techniques to handle the class imbalance of the data and then try fitting the logistic regression model again. Apart from these, we also want to train non-traditional simple neural networks for the federal prediction classification problem and LSTM (Long Short-Term Memory) networks for the time-series predictions.

- It has been a great learning experience for all of us working on this course project. The toxic release dataset provided is challenging and has given us real-world complexity handling sense. We have got good exposure to R and Python programming languages and different libraries.

Bibliography

1) Price, C. V., & Clawges, R. M. (1999). Digital data sets describing water use, toxic chemical releases, metropolitan areas, and population density of the conterminous United States. US Geological Survey Open-File Report, 99, 78.

2) Mastromonaco, R. (2015). Do environmental right-to-know laws affect markets? Capitalization of information in the toxic release inventory. Journal of Environmental Economics and Management, 71, 54-70.

3) https://www.epa.gov/toxics-release-inventory-tri-program/tri-basic-data-files-calendar-years-1987-2015

4) Lutz, M. (2001). Programming python. " O'Reilly Media, Inc.".

5) Fisher, M. J., & Marshall, A. P. (2009). Understanding descriptive statistics. Australian critical care, 22(2), 93-97.

6) https://northeastern.instructure.com/courses/97840/modules

7) Spector, P. (n.d.). Using t-tests in R | Department of Statistics. Https://Statistics.Berkeley.Edu/Computing/r-t-Tests

8) https://www.sheffield.ac.uk/polopoly_fs/1.536444!/file/MASH_2way_ANOVA_in_R.pdf

9) Bevans, R. (2020, December 14). *A step-by-step guide to linear regression in R*. Scribbr. https://www.scribbr.com/statistics/linear-regression-in-r/

10) *Logit Regression | R Data Analysis Examples*. (n.d.). Stats. https://stats.oarc.ucla.edu/r/dae/logit-regression/

11) *One-way ANOVA - An introduction to when you should run this test and the test hypothesis | Laerd Statistics*. (n.d.). Statistics. https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide.php