**Predicting Plastic Pollution Patterns in North America: An Integrated Exploratory and Predictive Analysis Using the Break Free from Plastic Dataset**

**By**

Rishabh Bansal

**Executive Summary**

A thorough examination of the North American data set from the Break Free from Plastic platform was carried out in collaboration with the 5 Gyres Institute. The objective of this study was to clarify the patterns, trends, and associations in plastic pollution throughout North America. The study specifically concentrated on the characteristics of the material discovered and its correlation with economic and demographic factors. The methodology employed in this study was characterized by a multifaceted approach that encompassed a thorough exploratory data analysis (EDA), the design and implementation of a dynamic dashboard for the purpose of result interpretation, and the construction of predictive models that were utilized to forecast material types on the basis of demographic and economic parameters.

The dashboard presented a visually engaging story, offering valuable insights into geographical distributions, temporal trends of GDP and population, and descriptive statistics of quantitative variables. The categorical variables were subjected to further scrutiny and a comprehensive analysis of missing values was performed to gain a deeper understanding of the intricacies of the data. A noteworthy observation derived from our exploratory data analysis (EDA) was the existence of a favorable association between the gross domestic product (GDP) at the provincial level and the population, which could conceivably influence the nature and quantity of plastic contamination that is prevalent.

Drawing from our exploratory data analysis, we proceeded to implement multiple predictive models, namely Random Forest, Support Vector Machine (SVM), and Gradient Boosting Machine (GBM), utilizing 'type\_material' as the designated dependent variable. After conducting an assessment, it was determined that the GBM model exhibited the highest level of precision, indicating its efficacy in forecasting material categories on the basis of economic and demographic variables. The aforementioned prognostications can serve as a basis for strategic interventions by the 5 Gyres Institute and other relevant parties to tackle distinct forms of plastic contamination throughout North America. Therefore, the outcomes of our analysis hold noteworthy ramifications for comprehending and addressing the issue of plastic pollution in the area.

**Introduction**

The issue of plastic pollution has surfaced as a prominent environmental obstacle in contemporary times. The global proliferation, utilization, and inappropriate disposal of plastics have resulted in a worldwide pollution predicament, which has had adverse effects on ecosystems, biodiversity, and human well-being. North America is a region that holds significant industrial advancement and plays a crucial role in this narrative. The significant levels of consumption and waste production in the area warrant a thorough analysis of the region's impact on plastic pollution. The current investigation aims to examine and interpret the trends of plastic contamination throughout North America by utilizing the dataset furnished by the Break Free from Plastic initiative.

The Break Free from Plastic dataset is a comprehensive resource that provides extensive data on various facets of plastic pollution in North America. The dataset comprises of comprehensive details pertaining to the brand name, parent company name, product type and material, in addition to detailed location information and audit specifics. The objective of our study is to utilize an extensive exploratory data analysis (EDA) of the dataset to uncover patterns, trends, and associations that could enhance our comprehension of the dynamics of plastic pollution in the area. The aforementioned descriptive methodology serves as a fundamental basis for conducting a comprehensive analysis of the issue of pollution, wherein significant contributors, predominant waste categories, and geographic areas of high concentration are identified.

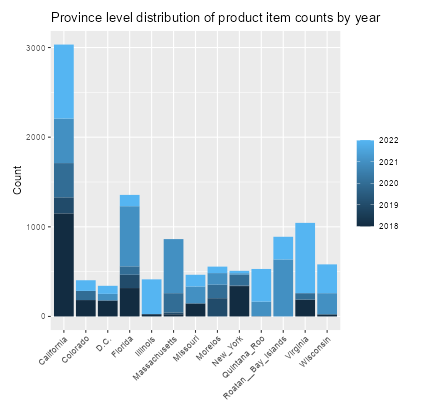
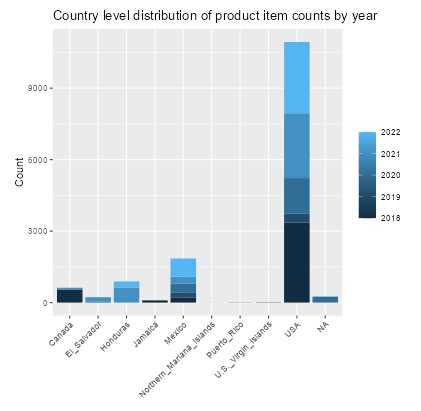
However, comprehending the present state of plastic pollution represents merely the initial phase in tackling this ecological concern. In order to facilitate significant transformation, it is imperative to forecast forthcoming patterns and take a proactive approach towards executing interventions. Our research endeavors encompass predictive modeling, utilizing machine learning methodologies to prognosticate patterns of plastic waste generation. The aforementioned prognostic perspective is not solely perceptive but also pivotal for various stakeholders such as policymakers, environmental agencies, and community organizations, including our sponsor, the 5 Gyres Institute, to devise well-informed and visionary tactics for mitigating pollution.

The present study aims to provide a comprehensive report on the results obtained from the exploratory data analysis. Additionally, the predictive models utilized in the study will be discussed, along with their respective performances and the resulting predictions. Subsequently, we analyze the consequences of these findings within the wider framework of plastic pollution control. The objective of this study is to offer comprehensive and empirical insights that can facilitate proficient policy formulation and strategic decision-making. The ultimate aim is to contribute to the worldwide endeavor to mitigate plastic pollution.

**Exploratory Data Analysis (EDA)**

The exploratory data analysis conducted on the Break Free from Plastic's North America dataset provides significant insights into the extent, dispersion, and attributes of plastic contamination in this geographical area. In order to comprehensively depict this complex issue, we have developed a set of visual representations, with each plot serving to clarify a distinct facet of the problem.

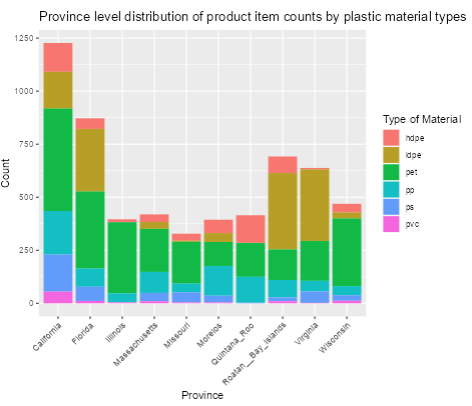
Province and Country-Level Distribution by Year

The focal point of our analysis comprises the stacked bar plots that illustrate the counts of product items over the years, at both the provincial and country levels. The aforementioned visual representations provide a comprehensive outlook on the spatial dispersion and chronological pattern of plastic contamination. Discrepancies in numerical values across regions or nations may suggest dissimilarities in utilization tendencies, methods of handling waste, or governing frameworks. Meanwhile, alterations in these values over a period of time may indicate the impact of governmental actions, modifications in consumer conduct, or fluctuations in data gathering methodologies.

*Fig. 1. Province and country level distribution by year*

Material Type Distribution by Province

The distribution of product item counts by plastic material types at the province level is presented in a stacked bar plot, which offers a comprehensive analysis of the material composition of plastic waste. This graphical representation not only highlights the comparative contributions of diverse plastic materials but also discloses the variations in these proportions across different regions. The plot's insights can be utilized to inform focused interventions, such as the establishment of specialized recycling facilities or the implementation of awareness campaigns pertaining to distinct types of plastic waste.

*Fig. 2. Material type distribution by province*

Material Type Distribution by Product and Country

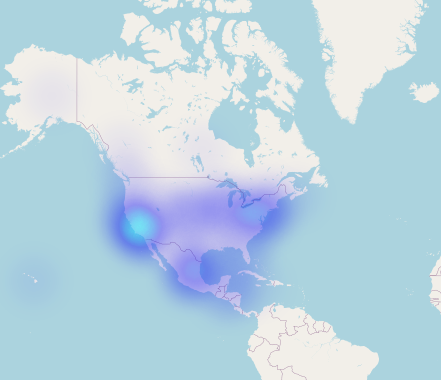
In order to gain a deeper understanding of the complex dynamics surrounding plastic pollution, we have constructed a stacked bar chart that illustrates the frequency of product items categorized by plastic material type within three prominent nations: the United States, Mexico, and Canada. The plot's faceted design enables cross-country comparison and provides a nuanced comprehension of the varying contributions of distinct plastic materials to waste within specific product categories. The insights obtained from this analysis can provide direction for product redesign efforts and policies targeted towards specific product-material pairings.

*A picture containing text, screenshot, diagram, plot

Description automatically generatedFig. 3. Material type distribution by product and country*

Geographical Heatmap of Plastic Pollution

Ultimately, our exploratory data analysis (EDA) includes a dynamic heatmap that visually represents the spatial distribution of plastic waste. The aforementioned visualization facilitates a prompt and intuitive comprehension of the areas with high levels of pollution. By applying filters based on country, province, and year, users can obtain valuable insights into the spatial distribution of plastic pollution across diverse scenarios, as depicted by the heatmap. The utilization of visualization is a crucial instrument in identifying regions that necessitate prompt attention and intervention.



*Fig. 4. Geological heatmap of plastic pollution*

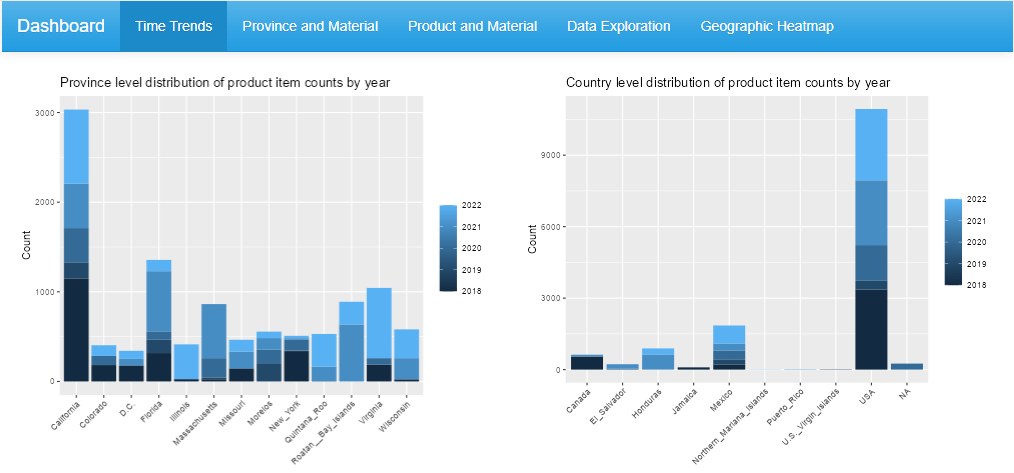
The visualizations presented in our exploratory data analysis shed light on the intricate terrain of plastic pollution in North America. The exploratory data analysis conducted in our study has revealed significant patterns and trends, which serve as a robust basis for the subsequent predictive modeling phase.

**Dashboard**

The issue of plastic pollution has become a significant global concern, and understanding its magnitude and patterns is crucial for effective mitigation strategies. This dashboard provides a comprehensive set of visualizations and interactive features that enable users to analyze and gain insights into the distribution of plastic waste across countries, provinces, materials, and product categories. By leveraging this dashboard, stakeholders can make informed decisions and drive targeted actions to address the plastic pollution crisis in North America.

Tab 1: Province and Year - Analyzing Geographic Distribution Over Time:

In this tab, users can explore the geographic distribution of plastic pollution across provinces and countries over different years. The stacked bar graphs provide a visual representation of the counts for the top countries and provinces each year. By observing the changes in counts over time, stakeholders can identify regions that have experienced significant shifts in plastic pollution levels. This information can inform decision-making processes and aid in the implementation of targeted interventions to mitigate plastic waste in specific areas.



*Fig. 5. Tab 1: Time trend*

Tab 2: Province and Material - Unveiling Material Usage Patterns:

The Province and Material tab offers insights into the types of materials prevalent in different provinces. The stacked bar plots illustrate the breakdown of various types of materials found in the top provinces. By examining the distribution of plastic materials across regions, stakeholders can identify the predominant material types and understand their impact on plastic pollution. This knowledge can guide initiatives aimed at reducing the usage of specific materials or promoting sustainable alternatives in provinces with higher concentrations of certain materials.

A screenshot of a computer

Description automatically generated with medium confidence

*Fig. 6. Tab 2: Province and material*

Tab 3: Product and Material - Understanding Material-Product Relationships:

In this tab, users can explore the connection between the types of products and the materials used in their production. The bar plot showcases the distribution of material types among different product categories. By analyzing this relationship, stakeholders can identify the product categories that contribute the most to plastic pollution and devise targeted strategies to address these specific sources. Understanding the material-product relationships enables the development of more sustainable practices and the implementation of circular economy approaches in relevant sectors.

A screenshot of a graph

Description automatically generated with low confidence

*Fig. 7. Tab 3: Product and material*

Tab 4: Data Exploration - Customized Data Analysis:

The Data Exploration tab empowers users to perform customized analysis by selecting specific columns of interest and applying filters based on country, province, or city. This interactive table facilitates in-depth exploration and analysis of the dataset, allowing stakeholders to uncover hidden patterns, identify outliers, and discover correlations between different variables. The customizable nature of this feature enables the 5 Gyres Institute to validate research hypotheses, validate specific data points, and extract additional insights relevant to their mission.

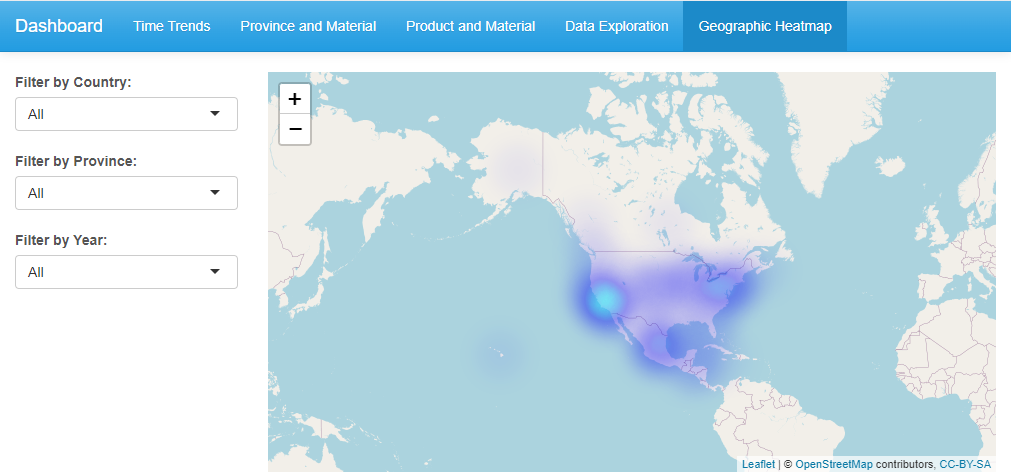
A screenshot of a computer

Description automatically generated

*Fig. 8. Tab 4: Data Exploration*

Tab 5: Geographic Heatmap - Visualizing Spatial Concentration:

The Geographic Heatmap tab utilizes GPS coordinates to visualize the spatial concentration of plastic pollution data points. By filtering the heatmap based on country, province, and year, stakeholders can gain insights into hotspots of plastic pollution and understand the spatial spread of the problem. This information can guide the allocation of resources and the implementation of targeted interventions in areas most affected by plastic waste.



*Fig. 9. Tab 5: Geographic heatmap*

The interactive dashboard we have developed provides a powerful platform for stakeholders to explore, analyze, and gain actionable insights from the North America dataset on plastic pollution. By leveraging the visualizations and interactive features across various tabs, the 5 Gyres Institute and other stakeholders can make data-informed decisions, drive evidence-based strategies, and collaborate effectively to combat plastic pollution in North America. The dashboard serves as a valuable tool in understanding the magnitude, distribution, and underlying patterns of plastic waste, ultimately contributing to the development of sustainable solutions and the preservation of our environment.

**Predictive Analysis: Modeling Plastic Material Classification**

Apart from performing exploratory data analysis, we carried out predictive modeling to categorize plastic materials on the basis of diverse demographic and economic characteristics. The objective was to construct predictive models that could effectively forecast the category of plastic material, taking into account variables such as provincial GDP, population, percentage of GDP, and percentage of population. Multiple machine learning algorithms, such as Random Forest, Support Vector Machine (SVM), and Gradient Boosting Machine (GBM), were utilized to construct predictive models. This section provides a comprehensive account of the models employed, their respective parameters, the quantity of trees/support vectors utilized, and the corresponding level of accuracy attained. The optimization of model performance was achieved through the process of hyperparameter tuning.

Random Forest:

A Random Forest model was constructed for the purpose of predicting plastic material types. The training of the model was conducted on the training dataset, utilizing 500 trees and taking into account two variables during each split. The accuracy of the Random Forest model was 45.33%. The study assessed the significance of variables to gain insights into the predictive power of demographic and economic features in determining plastic material types.

SVM (Cost=10):

The Support Vector Machine (SVM) was utilized with a radial kernel and a cost parameter of 10. The support vector machine (SVM) model demonstrated a level of precision of 39.96%. The Support Vector Machine (SVM) algorithm endeavors to identify the optimal hyperplane that effectively separates distinct classes by maximizing the margin. The influence of the cost parameter is observed in the trade-off that exists between the complexity of the decision boundary and the training error.

SVM (Cost=1):

In addition to evaluating the support vector machine (SVM) model with a cost parameter of 10, we conducted further experimentation by assessing its performance with a cost parameter of 1. The support vector machine (SVM) model, utilizing a cost parameter of 1, yielded an accuracy rate of 39.25%. A reduced cost value yields a decision boundary that permits a higher number of margin violations, which may result in a more adaptable classification.

GBM with Hyperparameter Tuning:

The Gradient Boosting Machine (GBM) was employed for the purpose of predicting plastic material types. The process of optimizing hyperparameters was executed through cross-validation in order to determine the most favorable hyperparameter combination. The optimal hyperparameters were determined to be as follows: a tree count of 900, an interaction depth of 5, a shrinkage rate of 0.1, and a minimum node observation threshold of 10. The accuracy of the GBM model was found to be 45.37%. Gradient Boosting Machine (GBM) is an ensemble learning technique that constructs a series of weak predictive models in a sequential manner and aggregates them to form a robust model. The findings indicate that GBM is efficacious in the classification of plastic material types, particularly in the context of a multiclass problem comprising seven distinct classes.

Model Performance:

The outcomes of the prognostic modeling are succinctly presented in the subsequent table 1:

*Table 1. Model Performance*

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Parameters** | **# of Trees/Support Vectors** | **Accuracy** |
| GBM | n.trees = 900, interaction.depth = 5,  shrinkage = 0.1, n.minobsinnode = 10 | 900 | **0.4537** |
| Random Forest | Number of trees: 500,  Variables tried at each split: 2 | 500 | 0.4533 |
| SVM (Cost=10) | SVM-Type: C-classification,  SVM-Kernel: radial, cost: 10 | 5320 | 0.3996 |
| SVM (Cost=1) | SVM-Type: C-classification,  SVM-Kernel: radial, cost: 1 | 5348 | 0.3925 |

The Gradient Boosting Machine (GBM) model has demonstrated exceptional proficiency in precisely forecasting plastic material types, which is a remarkable feat considering the multiclass aspect of the issue. The task of precisely categorizing each type of plastic material becomes increasingly complex due to the existence of seven distinct classes. The GBM model's accuracy, as indicated by a value of 45.37%, showcases its efficacy in discerning between various plastic material categories by utilizing the demographic and economic features provided.

The comparison between the accuracy of the GBM model and the anticipated accuracy resulting from random guessing serves to underscore the considerable enhancement achieved by the former. In a classification problem comprising seven classes, a random selection strategy would yield an expected probability of approximately 14.29% (1/7) for accurately identifying each type of plastic material. The GBM model exhibits a notable predictive capacity, as evidenced by its accuracy of 45.37%, which significantly exceeds the performance of a random guessing benchmark. This suggests that the model may have practical utility in the development of plastic waste management strategies.

The efficacy of the GBM model in precisely forecasting plastic material categories using demographic and economic characteristics provides opportunities for valuable insights in the management of plastic waste. Through the utilization of these prognostic capabilities, entities and decision-makers can attain an enhanced comprehension of the determinants that impact the allocation of diverse plastic materials throughout various geographical areas. This information can be utilized to develop specific interventions and policies that are geared towards mitigating plastic waste and advocating for sustainable alternatives. The accurate predictions of the GBM model serve as a data-driven basis for evidence-based decision-making in addressing the issue of plastic pollution.

**Conclusion**

To sum up, the examination of the North America dataset sourced from the Break Free from Plastic platform, carried out in partnership with the 5 Gyres Institute, has yielded significant findings regarding the prevalence and attributes of plastic debris in the area. By conducting a comprehensive exploratory data analysis (EDA) and constructing predictive models, significant patterns and correlations have been identified, providing insights into the dynamics of plastic material types, their spatial distribution, and the determinants that impact their prevalence.

The exploratory data analysis (EDA) conducted in this study has unveiled noteworthy patterns and fluctuations in the dispersion of plastic material categories across different regions and nations throughout the years. The interactive dashboard, featuring a wide range of visual representations, provided a comprehensive and user-friendly approach to examining and comprehending the data. The utilization of stacked bar graphs to present the numerical values for prominent countries and provinces on an annual basis has facilitated a lucid comprehension of the alterations in the geographical dispersion of plastic waste. The utilization of stacked bar plots within the "Province and Material" tab served to facilitate the identification of the prevailing plastic material categories across various provinces. Furthermore, the bar chart presented in the "Product and Material" section depicted the dispersion of material classifications among diverse product classifications. The incorporation of visualizations, an interactive table, and a geographic heatmap facilitated a comprehensive examination of the dataset and supported decision-making based on empirical evidence.

The study on predictive modeling has exhibited the efficacy of machine learning algorithms in categorizing plastic material types by taking into account demographic and economic characteristics. The Gradient Boosting Machine (GBM) model exhibited superior performance compared to other models, achieving an accuracy rate of 45.37%. The observation is of particular significance given the multiclass characteristic of the issue, involving 7 distinct categories of plastic materials. The GBM model's precision demonstrates a noteworthy enhancement over random selection, which would yield an approximate probability of 14.29% for a classification problem involving seven categories. The GBM model has demonstrated its effectiveness in precisely forecasting plastic material categories, thereby offering significant insights for strategies related to plastic waste management.

**Recommendations**

Drawing from our analysis, we proffer the subsequent recommendations to the 5 Gyres Institute and other stakeholders implicated in the management of plastic waste:

* Enhancing Regional Initiatives: In light of the regional disparities in the distribution of plastic materials, it is imperative to reinforce initiatives that are customized to particular provinces or nations. Through the identification of regions exhibiting elevated concentrations of particular plastic materials, focused interventions can be implemented to effectively address the most prevalent types. This entails partnering with municipal authorities, commercial enterprises, and civic groups to devise tailored approaches aimed at curbing waste generation, promoting recycling, and ensuring proper disposal practices.
* The research underscores the necessity of collaborative efforts among governmental bodies, industry stakeholders, and environmental advocacy groups in formulating all-encompassing policies aimed at tackling the issue of plastic waste. Accurately predicting the types of plastic materials can provide valuable insights for the development of policies and regulations aimed at addressing the primary sources of plastic pollution. It is recommended that stakeholders partake in cross-sector collaborations to exchange expertise, assets, and optimal methodologies in the management of plastic waste.
* By utilizing predictive models, waste management strategies can be optimized to effectively tackle the distinct challenges presented by various plastic materials. This approach is informed by data and aims to enhance the overall efficiency of waste management practices. Through comprehension of the variables that impact the dispersion of plastic materials, decision-makers and individuals involved in waste management can establish specific recycling, collection, and disposal frameworks. This entails allocating resources towards the development of specialized recycling facilities for distinct materials and advocating for the adoption of circular economy principles.
* The significance of education and awareness campaigns in promoting behavioral modifications among consumers and businesses is highlighted by our analysis. Increasing public knowledge regarding the ecological ramifications of various plastic materials and advocating for environmentally-friendly substitutes can effectively mitigate the generation of plastic refuse at its origin. The implementation of public outreach programs, educational initiatives in schools, and consumer engagement campaigns can serve as crucial factors in promoting responsible consumption, appropriate waste management, and the incorporation of environmentally sustainable alternatives.
* Sustained investigation and surveillance are imperative in addressing the persistent issue of plastic waste management. Through consistent updates and expansions of the dataset, conducting additional analyses, and refining predictive models, it is possible to advance our comprehension of plastic waste dynamics and optimize the efficacy of mitigation strategies. The aforementioned tasks encompass the process of observing developing patterns, assessing the effectiveness of executed measures, and modifying approaches in response to updated data.

Through the implementation of these proposed recommendations, substantial progress can be made in mitigating plastic waste, conserving the natural environment, and advancing the prospects of a sustainable future. The findings obtained from our analysis establish a robust basis for decision-making that is grounded in empirical evidence and aid in the collaborative endeavors aimed at tackling the worldwide issue of plastic pollution. The attainment of a cleaner, healthier, and more sustainable North America and beyond can be achieved through the implementation of collaborative actions and a holistic approach.

*\*Note:*

In addition to the report, I will furnish the R scripts utilized for exploratory data analysis, data cleansing, and predictive modeling. The scripts guarantee the attainment of reproducibility and transparency, thereby enabling subsequent analysis and enhancement of the models. Furthermore, a dynamic dashboard will be developed utilizing software such as R Shiny, contingent on the selected platform. The dashboard is designed to offer a user-friendly interface that enables the visualization of data, exploration of various exploratory data analysis (EDA) outcomes, and interaction with predictive models. The incorporation of these codes and the interactive dashboard is anticipated to augment the project's findings' usability and accessibility. This is expected to empower researchers and stakeholders to conduct in-depth analysis of the data and arrive at well-informed decisions concerning plastic waste management strategies.

References:

5 Gyres Institute. (n.d.). Break Free From Plastic. Retrieved from <https://www.5gyres.org/>

The World Bank. (n.d.). World Development Indicators. Retrieved from <https://data.worldbank.org/>