ALY6080: Integrated Experiential Learning

Module 12 XN Final Project: Final Deliverables

Academic Term: Spring 2023
Respected professor: Ajit Appari

Submitted BY,

Vaibhav Rajput

Abstract:

Plastic pollution poses a significant threat to marine life and human health. This report provides a comprehensive analysis of plastic waste in North America, drawing attention to the role of products and parent companies in contributing to this issue. Referring to the dataset comprising information on product attributes such as brand name, parent company, material type, and geographical location, this report engages in extensive data preparation followed by exploratory data analysis. Through graphical representations such as frequency distributions, bar graphs, word clouds, and scatter plots, insights into the distribution of plastic materials and their contributing sources are unveiled. A logistic regression model is employed to predict the presence of plastic waste based on product attributes, with a focus on product types. The results reveal patterns and associations that have implications for mitigating the environmental impact of plastic pollution. Additionally, interactive dashboards offer a visual presentation of the data, highlighting the contribution of parent companies to plastic waste and the geographical distribution of contaminants. The findings underscore the necessity for concerted efforts in reducing plastic pollution through awareness, policy advocacy, and sustainable alternatives.

Introduction:

In response to this global problem, numerous organizations have taken initiatives to address plastic pollution. One prominent organization is 5 Gyres, dedicated to reducing plastic pollution through research, education, and advocacy. 5 Gyres conducts scientific expeditions to study and document plastic pollution in the world's oceans, raising awareness about the issue and advocating for policies to reduce plastic waste.

Other similar organizations, such as The Ocean Cleanup, Plastic Oceans International, Surfrider Foundation, and Greenpeace, have also made significant contributions. The Ocean Cleanup focuses on developing advanced technologies to remove plastic debris from the ocean. Plastic Oceans International works towards promoting behavioral change and raising awareness about plastic pollution through film and media campaigns. The Surfrider Foundation focuses on protecting coastal areas from plastic pollution, while Greenpeace advocates for systemic changes to reduce plastic production and promote sustainable alternatives.

This report analyzes a dataset on products audited in North America to find patterns and insights, focusing on plastic waste. The dataset includes attributes like brand name, parent company, year of audit, product type, material type, and more.

The report begins with thorough data preparation, followed by exploratory data analysis techniques such as frequency distribution, correlation matrix, and graphical representations. Logistic regression is used to predict plastic waste based on product attributes, providing insights for mitigating environmental impacts. Interactive dashboards present visualizations of parent company contributions, temporal analysis of plastic contaminants, product and material distributions, and global company presence. The report aims to inform decision-making and raise awareness about plastic waste.

Methodology

Data Description

The dataset encompasses a wide range of products, offering valuable insights into the brand, parent company, material composition, and other relevant attributes. These variables will be analyzed to gain a comprehensive understanding of the audited products and their characteristics in order to support informed decision-making and strategic planning.

The data preparation process involved handling missing values and cleaning the data. The missing values in the categorical columns such as 'Brand Name', 'Product Type', 'Material Type', 'Layer', 'City', 'Province', and 'Specifics of Audit' were replaced with the string 'unknown'. Furthermore, the ‘Layer’ column had values labeled as 'unsure', which were also replaced with 'unknown' for consistency. In the 'Total Count' column, missing values were imputed by the mean value of the non-missing entries in that column. Another important step in data preparation involved standardizing the 'Province' column by converting the values to lowercase and replacing them with standardized abbreviations using a pre-defined mapping. Additionally, dummy variables were created for some categorical variables like 'Parent Company', 'Brand Name', 'Province', 'City', and 'Product Type'. This was done to enable the use of the categorical variables in the regression model. The 'Total Count' column was converted into a binary column named 'Plastic Waste', where values greater than 10 were marked as 1 (indicating the presence of plastic waste), and others were marked as 0.

The variables considered in this report include:

Brand Name: This refers to the name of the brand associated with a specific product.

Parent Company: This indicates the name of the parent company that owns the brand.

Year: This denotes the year in which the audit of the product was conducted.

Product Type: This represents the category or type of product, such as fishing gear, food packaging, household products, and others.

Material Type: This describes the material from which the product is made. The dataset includes materials such as High Density Polyethylene (HDPE), Low Density Polyethylene (LDPE), Polyethylene Terephthalate (PET), Polypropylene (PP), Polystyrene (PS), and Polyvinyl Chloride (PVC).

Layer: This category indicates the layer of the product, distinguishing between single-layer and multi-layer products.

City: This specifies the city where the product audit took place.

Province: This denotes the province or state where the product audit was conducted.

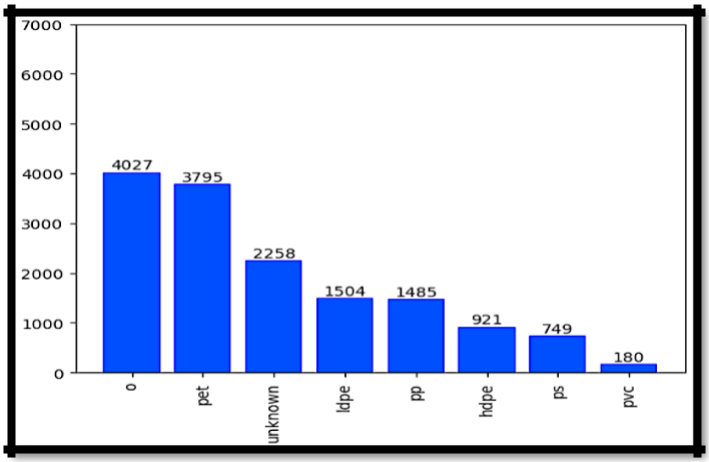
Specifics of Audit: This provides additional details about the audit conducted on the product.

Total Count: This represents the total number of items audited in the dataset.

For Prediction modeling, Logistic Regression was applied. This technique was used to model the probability of the presence of plastic waste based on selected product attributes ('Product Type'). Logistic Regression is commonly used for binary classification problems and has been widely applied in various fields including healthcare, finance, and marketing. For instance, in the context of environmental studies, logistic regression was used by Zhang, et al. (2019) in the article Logistic Regression Analysis of the Influencing Factors of Agricultural Plastic Film Residue (Journal of Physics: Conference Series, Volume 1176) to analyze factors influencing agricultural plastic film residue. In the logistic regression model used in this analysis, the target variable was 'Plastic Waste', and the independent variables included dummy variables derived from the 'Product Type' column. The model aimed at establishing a relationship between the type of product and the presence of plastic waste.

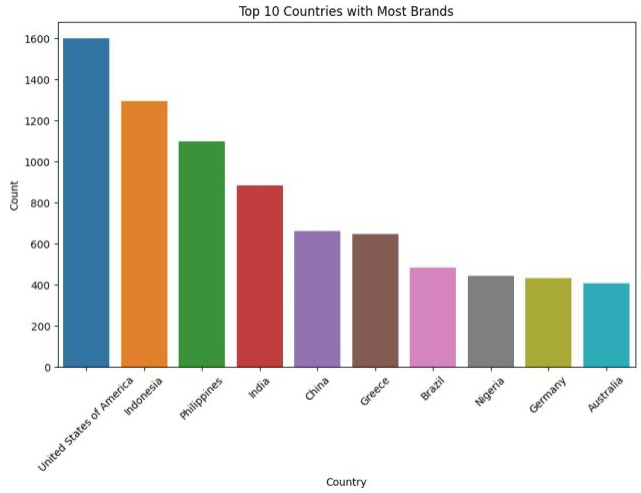
EDA Results and Discussion:

Bar graph of Frequency of Material Types in North America Dataset:



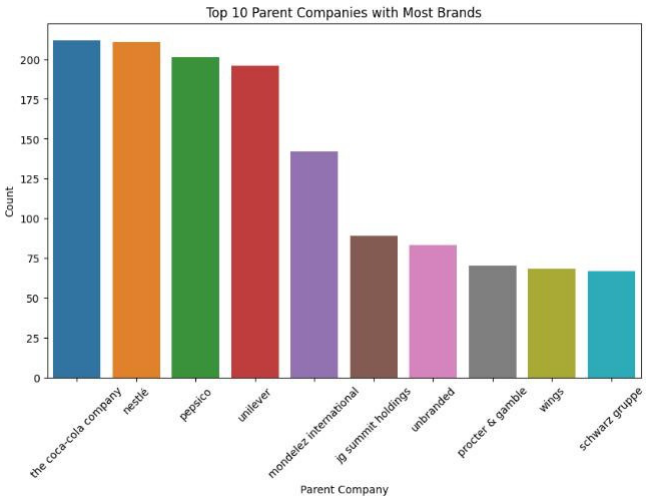
The above bar graph which represents the "Frequency of Material Types in North America", displays the occurrence of various material types in a dataset about North American products or waste. On the X-axis, material types are listed with vertically oriented labels, while the Y-axis shows the frequency. Each blue bar represents a material type, with its height indicating the number of occurrences and a label atop specifying the exact count. The graph is instrumental in understanding the distribution and prevalence of different materials in North America.

Top 10 Countries with the Most Brands:



This graph represents the top 10 countries with the highest number of brands. The x-axis represents the country names, and the y-axis represents the count of brands. The names of the countries are displayed at a 45-degree rotation for better readability. This plot helps to identify which countries have the most brands.

Top 10 Parent Companies with Most Brands:

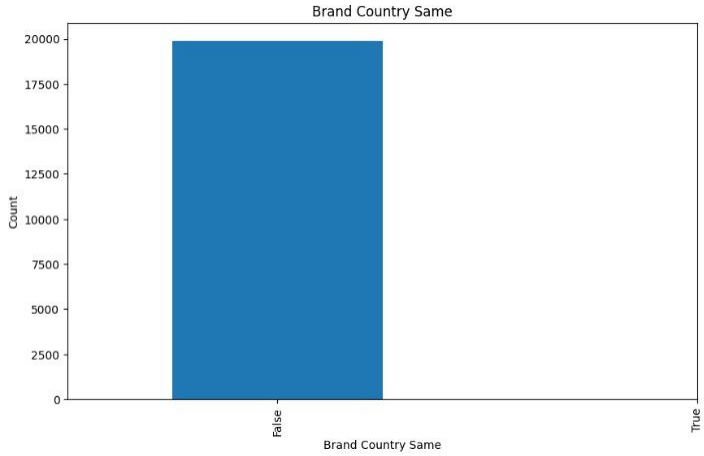


This graph represents the top 10 parent companies with the highest number of brands. The x-axis represents the names of the parent companies, and the y-axis represents the count of brands. This plot helps to identify which parent companies own the most brands.

Word Cloud of Brand Names:

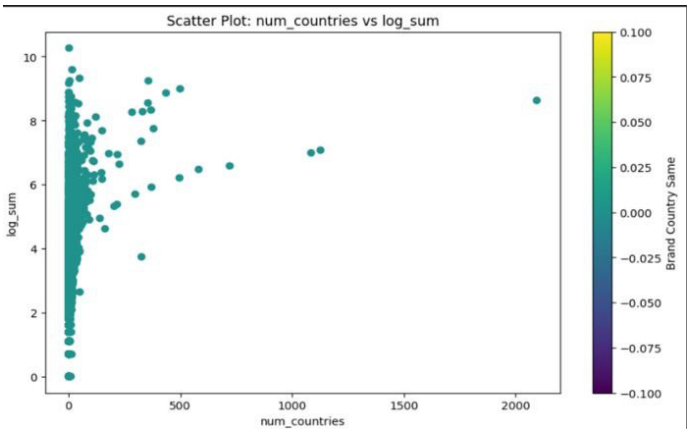
This word cloud graphically represents the frequency of brand names in the dataset. The size of each word in the image corresponds to its frequency. Larger words indicate higher frequency. This visualization helps to identify which brand names are more common in the dataset.

Brand Country Same Distribution:



This Bar graph displays the distribution of the 'brand_country_same' variable. The x-axis represents whether the brand name is the same as the country name ('True' or 'False'), and the y-axis represents the count. This plot provides insights into the number of records where the brand name is the same as the country name.

Scatter Plot of Number of Countries vs. Logarithm of Sum with Brand Country Same Coloring:



This scatter plot displays the relationship between the number of countries and the logarithm of the sum. Each point is colored based on whether the brand name is the same as the country name, with a color bar on the side indicating the scale. The x-axis represents the number of countries, and the y-axis represents the logarithm of the sum. This plot can be used to analyze the relationship between these two variables and how it relates to whether the brand name is the same as the country name.

Prediction Modeling Analysis and Discussion of Findings:

In this section, we analyze a dataset from North America to develop prediction models. The goal is to understand the relationships and patterns that can help in making predictions about various aspects related to the data, specifically focusing on the occurrence of plastic waste.

Logistic Regression Model, We employed Logistic Regression for predictive modeling, as it is well-suited for binary classification tasks. For our analysis, the target variable is "Plastic Waste Occurrence," which is a binary variable indicating whether the total count of plastic waste is above a certain threshold (10 in this case).

Feature Selection and Encoding:

The features selected for the predictive model encompass various product types. We transformed the categorical features into numerical representations. The features included are:

Product Type - Food Packaging

Product Type - Household Products

Product Type - Packaging Materials

Product Type - Personal Care Products

Product Type - Smoking-Related Materials

Model Estimation

The Logistic Regression model was estimated using the selected features to predict the presence of plastic waste. The results summary table includes valuable information about the coefficients, p-values, and other statistics for each feature.

Model Evaluation:

Logit Regression Results

Dep. Variable:	Plastic Waste	No. Observations:	14919
Model:	Logit	Df Residuals:	14913
Method:	MLE	Df Model:	5
Date:	Thu, 15 Jun 2023	Pseudo R-squ.:	0.003723
Time:	04:46:42	Log-Likelihood:	-4416.3
converged:	True	LL-Null:	-4432.8
Covariance Type:	nonrobust	LLR p-value:	3.748e-06

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.2594	0.043	-52.707	0.000	-2.343	-2.175
Product Type_food packaging	-0.1624	0.063	-2.598	0.009	-0.285	-0.040
Product Type_household products	-0.0386	0.120	-0.323	0.747	-0.273	0.196
Product Type_packaging materials	0.2350	0.206	1.143	0.253	-0.168	0.638
Product Type_personal care	-0.8985	0.217	-4.137	0.000	-1.324	-0.473
Product Type_smoking materials	0.2933	0.170	1.723	0.085	-0.040	0.627

1. The coefficient for each feature indicates the change in the log odds of the occurrence of plastic waste for a one-unit change in the corresponding product type.
2. The p-value indicates whether the coefficient is statistically significant at a given level of confidence. If the p-value is less than 0.05, it suggests that the coefficient is statistically significant.

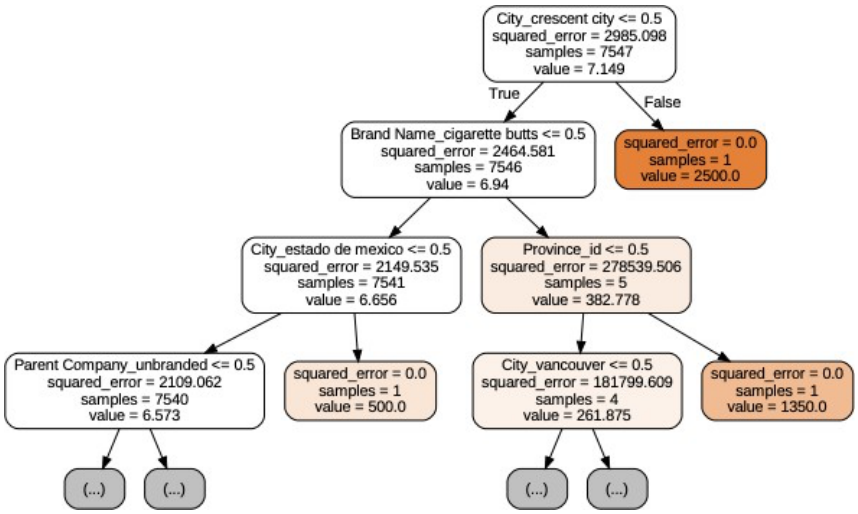
Random Forest: Predicting Plastic Waste

The random Forest model was developed for predicting plastic waste in the context of the project. The Random Forest algorithm is a powerful ensemble learning method that combines multiple decision trees to make accurate predictions. In this project, the Random Forest model has been trained on a dataset containing various features related to plastic waste to predict the total count of plastic waste items.

Model Description

The Random Forest model is built using the sci-kit-learn library in Python. It consists of an ensemble of decision trees, where each tree is trained on a random subset of the training data. The key steps involved in training and using the Random Forest model are as follows:

Data Preparation: The dataset used to train the model contains information about plastic waste, including features such as brand name, parent company, year, product type, material type, layer, city, province, and specifics of the audit. The target variable is the total count of plastic waste items. Before training the model, the dataset is split into features (X) and the target variable (y). Categorical variables are encoded using one-hot encoding to convert them into numerical format.



Training the Model: The Random Forest model is initialized with hyperparameters such as the number of decision trees (n_estimators) and a random seed for reproducibility. The model is then trained on the training data using the fit() function, which fits each decision tree to a random subset of the training data.

Making Predictions: After training, the model can be used to make predictions on new, unseen data. The model takes the features of the new data and predicts the total count of plastic waste items using the predict() function.

Model Evaluation: The performance of the Random Forest model is evaluated using mean squared error (MSE), which measures the average squared difference between the actual and predicted values. A lower MSE value indicates better performance.

Model Results and Interpretation

The trained Random Forest model provides insights into the relationship between the features and the total count of plastic waste items. It captures complex interactions and nonlinear relationships that may exist in the data. The model's performance can be assessed using the mean squared error (MSE) value, which quantifies the average prediction error. Additionally, the Random Forest model allows for feature importance analysis. By analyzing the importance scores assigned to each feature by the model, we can identify the most influential features in predicting plastic waste. This information can be used to understand the key factors contributing to plastic waste and guide targeted interventions or policy decisions.

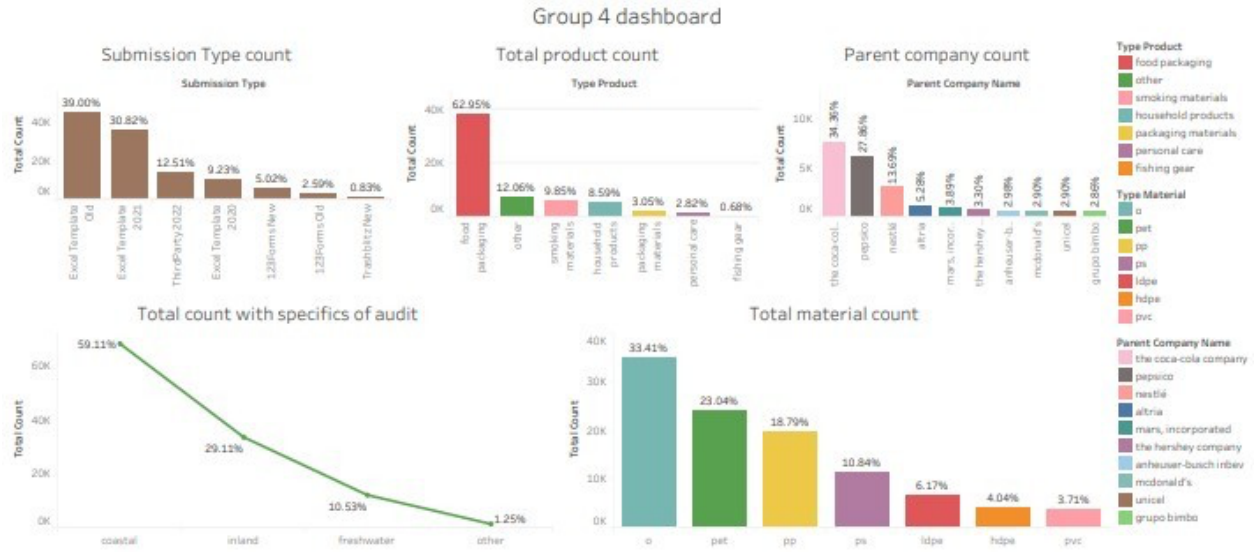
Visualizing the Random Forest

To gain further understanding of the Random Forest model, one can visualize individual decision trees within the forest. By visualizing the decision trees, we can observe the splitting criteria, feature importance, and decision rules learned by each tree. This can be accomplished using the graphviz library, which generates a visual representation of a decision tree.

Conclusion

The Random Forest model developed for predicting plastic waste in this project leverages the power of ensemble learning to provide accurate predictions. By training on a dataset containing various features related to plastic waste, the model can capture complex relationships and make informed predictions about the total count of plastic waste items. The model's performance can be assessed using metrics such as mean squared error (MSE), and feature importance analysis can provide insights into the factors influencing plastic waste. Visualizing individual decision trees within the Random Forest can enhance our understanding of the decision-making process

Dashboard 1:



Graph explanation for Parent Company Count:

Upon analyzing the data, Coca-Cola and PepsiCo emerge as the leading parent companies contributing to the generation of multi-layer plastic waste. It is imperative for Coca-Cola to recognize that the precision of this inference is contingent upon the integrity and comprehensiveness of the data examined. Moreover, it is essential to evaluate these companies' environmental policies and initiatives in the context of their role in global plastic pollution. A more holistic understanding of these findings and their ramifications necessitates further investigation and analysis.

Graph explanation about Total Count and Time Spent with Specifics of Audit:

This graph presents an aggregated analysis of plastic contaminants, categorized by distinct audit parameters. Coastal areas are the predominant locations where plastic pollutants were identified, comprising approximately 58.87% of the total, with 68,110 instances recorded. The second major contributor is inland areas, which represent roughly 29.02% of the total, with a count of 33,548. Freshwater ecosystems constitute around 10.47% of the total, with 12,139 instances. Other regions contribute minimally, approximately 1.63%, with 1,435 instances. The data implies that plastic

contaminants are more prevalent in coastal and inland areas. Furthermore, a temporal analysis reveals that coastal regions are particularly affected over time.

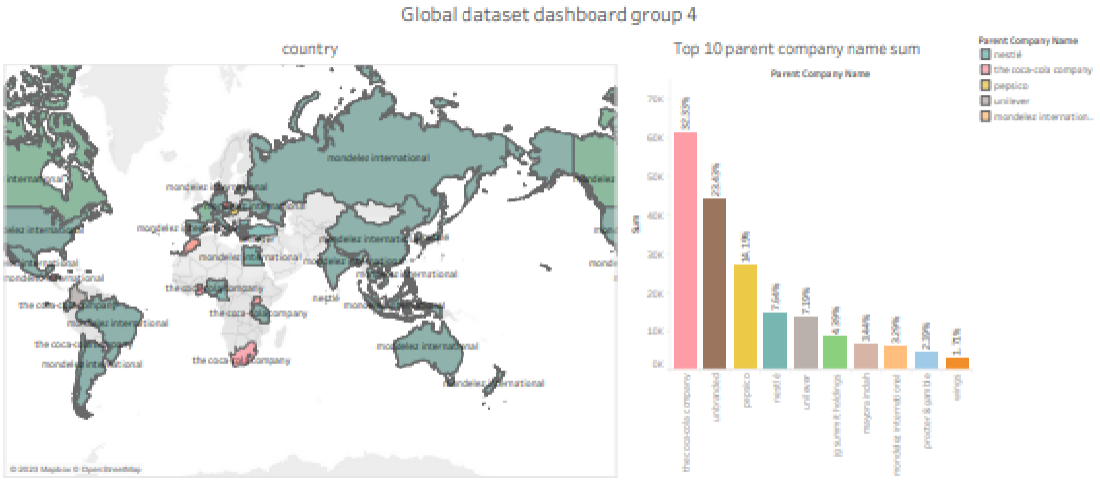
Graph explanation about Total Product Count:

This graph details the distribution of product types identified among the plastic contaminants, both in terms of numbers and percentages. A significant portion of the pollutants, about 47.45%, could not be traced back to a specific origin. Food packaging materials constituted the second-largest category, representing approximately 33.11% of the total with 38,272 instances. The category labeled as “other” was the third most common, encompassing about 6.3% of the total. Subsequent categories include smoking materials, household products, packaging materials, personal care products, and fishing gear.

Graph explanation about Total Material Count:

This graph provides a comprehensive overview of the various types of plastic materials identified within the contaminants, detailing their numbers and percentages. The material designated as “O” was the most prevalent, accounting for 30.27% of the total with 35,008 instances. PET plastic ranked as the second most common material, constituting approximately 20.89% of the total with 24,145 instances. PP plastic followed as the third most frequent material, representing 16.99% of the total. The remaining materials, including PS, unknown types, LDPE, HDPE, and PVC, are presented in descending order of prevalence.

Dashboard 2:



Geographical Map Description

The geographical map displayed at the outset provides a visual representation of the global distribution of companies in different countries. By utilizing the available filters, one can observe that Nestlé holds a prominent presence in a majority of these countries, suggesting a wide geographical reach for this company. The map employs different color codes or markers to indicate the density or count of companies in each country, with Nestlé seemingly dominating the representation.

Bar Chart Description

Following the geographical map, there is a bar chart that portrays the total number of entities associated with the top ten companies. It is interesting to note that despite Nestlé's extensive geographical presence, as shown in the preceding map, Coca-Cola outstrips other companies in terms of raw quantity. Coca-Cola boasts an impressive tally of 61,297 entities. The bar chart likely uses vertical bars with varying lengths to represent the counts of entities for each company. Coca-Cola's bar would be the tallest, reflecting its leading position among the top ten companies in terms of the number of entities.

Recommendations:

Based on the analysis and findings from the dataset on products audited in North America focusing on plastic waste, the following recommendations are made:

1. Raising Awareness on Plastic Waste: Given that food packaging materials constitute a significant portion of plastic contaminants, there is a need for consumer awareness campaigns to educate the public on the detrimental effects of plastic waste, especially in coastal areas where pollution is most prevalent.
2. Encourage Responsible Packaging: Coca-Cola and PepsiCo are the leading parent companies contributing to the generation of plastic waste. These companies should be encouraged to adopt more sustainable packaging alternatives and develop initiatives for recycling and reducing plastic waste.
3. Strengthening Waste Management Systems: In light of the high number of plastic contaminants in coastal and inland areas, strengthening waste management systems in these areas is crucial. This could include improving waste collection and recycling facilities, and adopting more sustainable waste disposal practices.

References:

1. 5 Gyres Institute. (n.d.). Fighting plastic pollution. Retrieved from https://www.5gyres.org/
2. The Ocean Cleanup. (n.d.). The largest cleanup in history. Retrieved from https://theoceancleanup.com/
3. Plastic Oceans International. (n.d.). Rethink plastic. Retrieved from https://plasticoceans.org/
4. Surfrider Foundation. (n.d.). Protect and enjoy. Retrieved from https://www.surfrider.org/
5. Greenpeace. (n.d.). Fighting for a greener, healthier world. Retrieved from https://www.greenpeace.org/international/
6. Zhang, et al. (2019). Logistic Regression Analysis of the Influencing Factors of Agricultural Plastic Film Residue. Journal of Physics: Conference Series, Volume 1176. Retrieved from https://iopscience.iop.org/article/10.1088/1742-6596/1176/6/062026/pdf
7. World Wildlife Fund. (2018). Solving Plastic Pollution Through Accountability. Retrieved from https://www.worldwildlife.org/stories/fighting-plastic-pollution-through-accountability

Appendix A: Additional Information on Research Organizations and Initiatives

This appendix provides additional information on organizations that have been actively involved in combating plastic pollution, as mentioned in the introduction of the report.

1. 5 Gyres
    - Website: 5 Gyres
    - Founded: 2009
    - Summary: 5 Gyres is a non-profit organization dedicated to reducing plastic pollution through research, education, and advocacy. The organization conducts scientific expeditions to study and document plastic pollution in the world's oceans. Its findings are utilized to raise awareness and advocate for policies to reduce plastic waste.
2. The Ocean Cleanup
    - Website: The Ocean Cleanup
    - Founded: 2013
    - Summary: The Ocean Cleanup aims to develop advanced technologies to extract, prevent, and intercept plastic pollution. It focuses on removing plastic debris from the oceans, rivers, and other water bodies. The organization's main project involves the deployment of passive cleanup systems in the Great Pacific Garbage Patch.
3. Plastic Oceans International
    - Website: Plastic Oceans International
    - Founded: 2009
    - Summary: Plastic Oceans International focuses on promoting behavioral changes and raising awareness about plastic pollution through film and media campaigns. The organization aims to educate the global audience about the impact of plastic pollution on the environment and human health.
4. Surfrider Foundation
    - Website: Surfrider Foundation

- Founded: 1984
- Summary: Surfrider Foundation is an environmental organization that protects and preserves the world's oceans and beaches. It emphasizes reducing plastic pollution in coastal areas through beach cleanups, educational programs, and advocacy efforts.

5. Greenpeace
    - Website: Greenpeace
    - Founded: 1971
    - Summary: Greenpeace is a global non-profit organization that focuses on environmental issues such as climate change, deforestation, overfishing, commercial whaling, and anti-nuclear issues. It advocates for systemic changes to reduce plastic production and promote sustainable alternatives.

These organizations play a significant role in raising awareness, conducting research, and taking practical steps to address plastic pollution. By engaging in advocacy efforts and promoting education, these organizations are positively impacting the fight against plastic pollution.

This appendix can be referenced for more detailed information on the organizations that are taking initiatives to address the issue of plastic pollution as mentioned in the introduction of the report.