ANALYZING PLASTIC MARINE DEBRIS TO IDENTIFY & PINPOINT SOURCES OF MARINE POLLUTION

Andish Bagheri, Blake Bensman, Hannah Stansfield, & Alexios Prodanas

Abstract:

Water pollution has noticeably become a threat toward humans and marine life, ultimately impacting the environment and ecosystem. Marine debris cleanup operation companies use different methods and strategies to remove pollution from the waters. It has been useful for companies to develop strategies and use analytics to aid in the cleanup process of marine debris. This paper will analyze the possible sources of marine debris and the effect of how each independent variable is associated with particular debris items. The overall effectiveness of our business intelligence solution will be presented using various modeling techniques, in order to properly consult clients that wish to effectively clean up marine debris. Furthermore, this paper will evaluate which type of marine debris is more common throughout the polluted waters.

Keywords: Marine, debris, hard plastics, soft plastics, ecosystem, cleanup, pollution, analytics, beaches

1. Introduction

Plastic marine debris is a growing problem for our world's oceans. Each year, 8 million metric tons of plastic end up in the ocean, collectively affecting the health of marine life and human communities ("11 Facts"). Marine debris is found across the world in various locations, such as beaches, oceans, and others. This problem has a collective impact on our human and natural world ecosystems. Marine debris cannot biodegrade normally like plants and accumulates in the water. As a result, marine species are ingesting this debris which has the potential to end up in our food systems, which can pose negative effects to our health. This provides an opportunity for businesses to tackle the marine debris problem and provide data-driven insights on how we can solve such a problem.

This report highlights the results from the consulting company, aBIT of Plastic Reduction, which analyzes large data related to debris found in oceans and waterways. The company then uses its findings to help businesses and communities reduce waste and prevent future oceanic pollution. It also specializes in taking datasets and using statistical modeling methods to analyze plastic marine debris to identify and pinpoint sources of marine pollution. As the premier data analysis organization, the company is interested in targeting marine debris in all forms and where they tend to be accumulating most prevalently.

We have been tasked by the company to analyze data collected by the NOAA on quantities of debris found in oceans in and around the US, Canada, Palau, and Ecuador. Reliance on this data and our findings will create a cleaner world that will provide healthier ecosystems and provide data-driven suggestions pertaining to which marine debris items should be targeted for cleanups, sustainable replacements, and pollution reduction efforts for businesses and communities.

Additionally, we feel our findings could be useful from an economic perspective for businesses. For example, we could consult businesses within our areas of study on our findings. These businesses could use our analyses to mitigate materials within their business models or manufacturing efforts to align with EPA pollution regulations. This could help businesses avoid or reduce fees and penalties associated with being charged by governmental entities for dumping and other forms of wasteful practices.

1.1 Background

Many researchers around the world have explored the topic of marine debris and its collective impact on our world's oceans. A plethora of individuals and organizations are looking for pathways to capture the marine debris and discover ways to incrementally decrease it.

In a study published by CSIRO Oceans and Atmosphere (2018), they provide suggestions on improving marine debris monitoring tools within the United States using a variety of statistical modeling techniques to analyze location-specific ocean debris patterns. Marine debris hotspots along shorelines of the United States are highlighted, citing the data collection NOAA Marine Debris Program's Marine Debris Monitoring and Assessment Project (MDMAP, 2020) and the Ocean Conservancy's International Coastal Cleanup dataset (CSIRO Oceans and Atmosphere, 2018).

Roosevelt et al. (2013) documents the spatial and seasonal patterns of marine debris, researchers in Monterey Bay, CA using sampling and surveying techniques to quantify the impact of marine debris beach pollution. Baselines for types of litter and their abundance were explored and analyzed, to provide justification for local legislation on banning items such as Styrofoam.

Morocco has a reputation for its beaches among locals and those visiting. Abundance, composition, and sources of marine debris are explored in this paper, in order to gain insights on ocean beach pollution in the Morocco Mediterranean highlighted in the study conducted by Mghil et al. (2020).

Taking a “dive” into the world of marine plastics and their sources, the study seeks to give a general overview of marine plastics impacts on both human and natural world communities. The authors, Van Truong et al. (2019), detail marine waste totals from the coasts of Vietnam, citing strategies for minimizing and managing plastic waste coming from ships.

1.2 Objectives, Limitations, Challenges and Modeling of Data

Our company is committed to thoroughly analyzing the marine debris data, in order to give our client direction on which marine debris to target through their ocean/beach clean up activities. To achieve these goals, we plan on targeting specific locations, looking at different weather patterns, and seasons, which we feel will help us gain insight on what communities or/and businesses to target for cleanup. We see this as a valuable business intelligence solution for mitigating marine debris across various water sources.

Like all data, there are some limitations with our marine debris study. Limitations in our study include zero values for location data, the results are taken from a relatively small sample size when compared to the rest of the world’s area and populations, and normalization of our dataset’s marine debris numerical variables for categorization purposes. This could be a limitation for our overall analysis because we used broader categories, in order to encompass more types of marine debris. In normalization of numerical variables, we found that multiple attributes that did not fit into a specific category and contained too many zero values for their own columns resulted in merging these into two separate miscellaneous columns.

This data provides an opportunity to use a wide array of models to analyze trends and significance. First off, we could use multiple regression modeling techniques to find relationships between the continuous variables (i.e. hard plastics, soft plastic, etc.) Second, principal component analysis will allow us to analyze variation of types of plastics found at specific locations and summarize strong patterns across the entire dataset. Third, we plan to use clustering in order to group similar plastics and other relevant materials into clusters (what type of marine debris is found in more or less occurrence).

2. Dataset Introduction

The initial dataset was obtained from Kaggle, “NASA Project; Marine Debris Machine Learning NASA Project; Plastic Marine Debris Classification-Machine Learning Software” (2021). The dataset was sourced from the NOAA (National Oceanic and Atmospheric Administration) Marine Debris Program (2017), in an effort to collect information about marine debris throughout the globe. The Kaggle dataset author provided an assortment of parameters (date, country, etc) using a variety of numerical variables (plastic type, metal type, etc.). In all, the data seeks to explore the irregularity of various marine debris data and its proportions for specific years and countries.

Our dataset contains 1,186 rows of data with 35 attributes, 12 of those being continuous variables to be used in our analysis. The other 23 attributes contain information about time and date; weather patterns; location data, including latitude and longitude, country, state, county, and shoreline; and season.

In Table 2.1, we present the description of our marine debris dataset, specifically indicating the variables as categorical or numerical. In addition, we describe the variable type and whether or not the data can be a dependent variable. This will be important for recognizing which variables to compare with various modeling techniques and summarize marine debris-related conclusions for our clients.

Table 2.1: Attributes Descriptions and Variable Types
N= Nominal C=Continuous I= Independent D=Dependent

Name	Description	N & C	I/D
Organization	Organization responsible for data collection.	N	I
Date	Date data was collected	N	I
Survey_Year	Number of years the area was surveyed for debris.	N	I
Country	Country data was collected	N	I
State	State data was collected	N	I
County	County data was collected	N	I
Shoreline_Name	Shoreline debris data was collected from	N	I
Latitude_Start	Latitude of location	C	D
Longitude_Start	Longitude of location	C	D
Latitude_End	Latitude of location	C	D
Longitude_End	Longitude of location	C	D
Start_Time	Start time of collecting samples	N	I
End_Time	End time of collecting samples	N	I
Time_of_Low_Tide	Time low tide was observed.	N	D
Database_Season	Season data was collected	N	D
Days_since_last_survey	Amount of days since last survey	N	I
Storm Activity	Storms observed during location survey and debris collection.	N	D
Current_Weather	Weather observed during data collection.	N	D
Number_of_person_Assisting	How many people aided in data collection	C	I
Large_Items	If the item collected was large or not	N	I
Debris_Behind_Back_Barrier	Amount of debris found behind the back barrier.	N	D
Hard_Plastics_Total	Amount of hard plastics found	C	I
Hard_Plastics_Total_Flux	Amount of micro particles of hard plastics found	C	D

Soft_Plastics_Total	**Amount of soft plastics found**	**C**	**D**
Soft_Plastics_Total_Flux	**Amount of Micro particles of soft plastic found**	**C**	**D**
Food_Items_Total	**Amount of food items found**	**C**	**D**
Food_Items_Total_Flux	**Amount of Micro particles of food items found**	**C**	**D**
Bottles_Total	**Amount of bottles found**	**C**	**D**
Bottles_Total_Flux	**Amount of Micro particles of bottles found**	**C**	**D**
Misc_Hard_Goods_Total	**Amount of hard miscellaneous items found.**	**C**	**D**
Misc_Hard_Goods_Total_Flux	**Amount of Micro particles of hard miscellaneous items found.**	**C**	**D**
Misc_Soft_Goods	**Amount of soft miscellaneous items found**	**C**	**D**
Misc_Soft_Goods_Flux	**Amount of Micro particles of soft miscellaneous items found.**	**C**	**D**
Total_Debris	**Number of total debris pollution**	**C**	**D**
Total_Debris_Flux	**Micro particles of total debris pollution**	**C**	**D**

2.1. Overview of Dataset Variables

We have chosen a few dependent variables to analyze for our client. Dependent variables will be an important focus for our analysis because they will show a causal relationship between the marine debris instances and their association(s) with certain conditions (dependent variables). Below we detail specific instances where independent variables will have the potential to impact or influence the marine debris variables.

We believe that depending on the season could have a big impact on the results and the accuracy of our results. For example, during the summer more people travel to the beach which could result in more pollution in the waters.

Since the storm can bring the pollution from all over the places to the water, we believe the activity of the storm will affect the number of debris pollution in the water. We will try to find a correlation between the storm activity and the number of pollution found during that time.

All these variables will help us to find connections between the different types of items collected and which types have the higher or lower occurrence. Additionally, we can get more insights on the factors that affect the increase or decrease of the pollution.

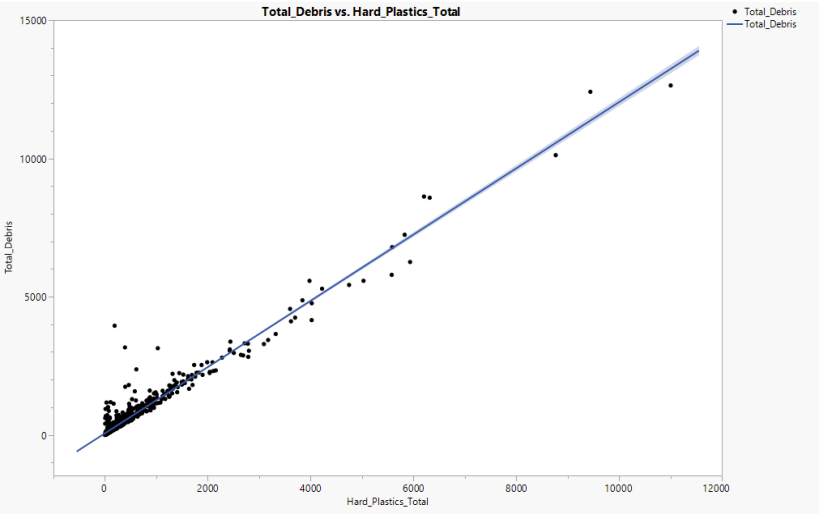
3. Data Exploration

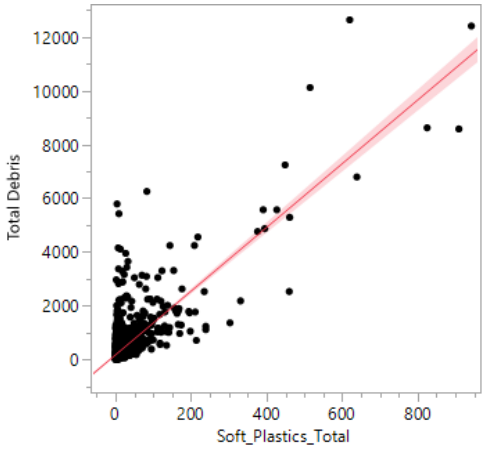
To successfully analyze the marine debris dataset, we have conducted a number of analyses to assist our client with pinpointing the types and locations of marine debris. In this section, we will perform initial tests such as scatterplot matrices, multiple regression, and one-way ANOVA. These tests will help determine which of our variables are significantly related and will show what needs to be investigated further. The firm does not want to spend extra time and money on variables that show little to no relation, so these initial tests are important.

3.1 Scatterplot Matrix Analysis

To begin, we analyzed the relationship between Total_Debris (dependent) and all other forms of debris (independent). Analyzing the pairwise correlations depicted in Figure 3.1 and Figure 2 in Appendix A, we noted significant relationships between variables that had a correlation of ~0.50, of which we had 16 relationships. This is significant for our business intelligence problem because we will be able to map out which correlations are the strongest to focus on when pinpointing marine debris types for client ocean cleanups. We observed that Total_Debris and Hard_Plastics_Total have the highest correlation value of 0.9775. This could tell us that Hard_Plastics are found more frequently and in higher amounts than other types of debris. The second highest correlation is between Soft_Plastics_Total and Total_Debris with a correlation of 0.8052. Similarly, this can tell us that Soft_Plastics_Total are also prevalent in the Total_Debris found in the examined locations. These two observations can be confirmed from the scatterplot matrix found in Figures 1 and 3 in Appendix A since the Hard_Plastics_Total and Soft_Plastics_Total have strong positive relationships with Total_Debris. Positive relationships are indicated with variables having narrow ellipses with linear positive correlations.

Figure 3.1: "Total Debris Scatter Plot vs Hard Plastics" and Total Debris Scatter Plot vs Soft Plastics"





3.2 Multiple Regression Analysis

Next, a multiple regression analysis was performed. Using the same variables as performed for the Scatterplot Matrix, As the R-Squared statistic is 0.231255, we can measure how variations of a dependent and independent variable(s) fit in a regression model. Here we can see that our dataset is able to explain the 23.12% of variation in our model. This is important for us to understand because each marine debris item is related, yet unrelated to each other. Meaning that, the lower RSquare value shows that marine debris will accumulate in the ocean and each marine debris item is independent from each other within the Total_Debris amount.

We use an Effect Summary Figure depicted in Figure 3.2 to find significant data based on their P-Values and LogWorth values in the model. Therefore, for the purpose of this project, we are focusing on: Hard_Plastics_Total with a P-value of 0.0 and the Longworth of 57.944; Food_Total_Flux with the P-value of 0.00109 and LogWorth of 2.964; Misc_Hard_Goods_Total_Flux with the P-value 0.0049 and LogWorth of 2.304. This regression model can tell us that these three sources are responsible for the highest number of micro pollutants in the ocean. Our company advises the organization to apply more regulation on the production of these specific items since micro pollution is harder to track and collect. As a result, it is harmful to the aquatic ecosystem including animals and plants as well as the human population since it can have negative effects on our health after consumption of the polluted seafood.

Since our VIF values for all of the parameters are between 1 and 5, we are confident there is no significant multicollinearity. This can help us understand variation in the dependent variable, which can give us more insights about the relationship between these parameters.

Figure 3.2: Multiple regression between Total_Debris and all flux of each flux category of marine debris

Effect Summary

Source	LogWorth	PValue
Hard_Plastics_Total_Flux	57.944	0.00000
Food_Items_Total_Flux	2.964	0.00109
Misc_Hard_Goods_Total_Flux	2.304	0.00496
Misc_Soft_Goods_Total_Flux	0.701	0.19928
Bottles_Total_Flux	0.402	0.39660
Soft_Plastics_Total_Flux	0.175	0.66817

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	1139	982171915	862311	1594.509
Pure Error	40	21632	541	Prob > F
Total Error	1179	982193547		<.0001*
				Max RSq
				1.0000

Summary of Fit

RSquare	0.231255
RSquare Adj	0.227343
Root Mean Square Error	912.7286
Mean of Response	469.9359
Observations (or Sum Wgts)	1186

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	295465001	49244167	59.1114
Error	1179	982193547	833073.41	Prob > F
C. Total	1185	1277658547		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Std Beta	VIF
Intercept	1509.4329	95.98971	15.72	<.0001*	0	.
Hard_Plastics_Total_Flux	-419.2929	24.53493	-17.09	<.0001*	-0.46924	1.156247
Soft_Plastics_Total_Flux	6.4637864	15.07539	0.43	0.6682	0.011995	1.2002235
Food_Items_Total_Flux	46.865865	14.31038	3.27	0.0011*	0.093823	1.2587571
Bottles_Total_Flux	-14.26791	16.82499	-0.85	0.3966	-0.02458	1.2889914
Misc_Hard_Goods_Total_Flux	42.69364	15.16802	2.81	0.0050*	0.079589	1.2262297
Misc_Soft_Goods_Total_Flux	18.565819	14.4556	1.28	0.1993	0.036242	1.2212313

In Figure 3.3, we used the Effect Summary Figure to narrow down our data and remove non-significant components. However, by removing the non-significant components, we received a lower RSquare value of 0.229751. This could imply that we should not prefer the model with the reduced components since the model is not able to give us a better explanation of our analysis and it provides us with less accurate results.

Figure 3.3: Effective Summary for Total Debris and Flux of Specific Marine Items

Effect Summary

Source	LogWorth	PValue
Hard_Plastics_Total_Flux	66.318	0.00000
Food_Items_Total_Flux	3.584	0.00026
Misc_Hard_Goods_Total_Flux	2.898	0.00127

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	1116	984082972	881795	1796.415
Pure Error	66	32397	491	Prob > F
Total Error	1182	984115369		<.0001*
				Max RSq
				1.0000

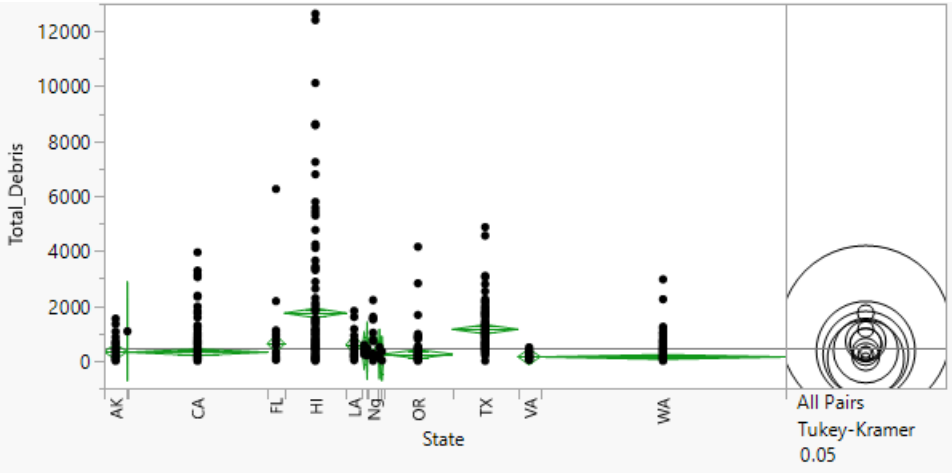
Summary of Fit

RSquare	0.229751
RSquare Adj	0.227796
Root Mean Square Error	912.4609
Mean of Response	469.9359
Observations (or Sum Wgts)	1186

3.4 ANOVA Analysis

In the next test we looked at the state and mean of total debris. The questions we strive to answer with this analysis are: Do the states have the same mean results? Do any of the groups have means that are different or not? We conducted a multiple-comparison test. Using "Means ANOVA," diamond shapes are created on the plot. The middle of the diamonds are the mean of the total debris, while the upper and lower shorter lines are the 95% confidence intervals and the size is the number of points.

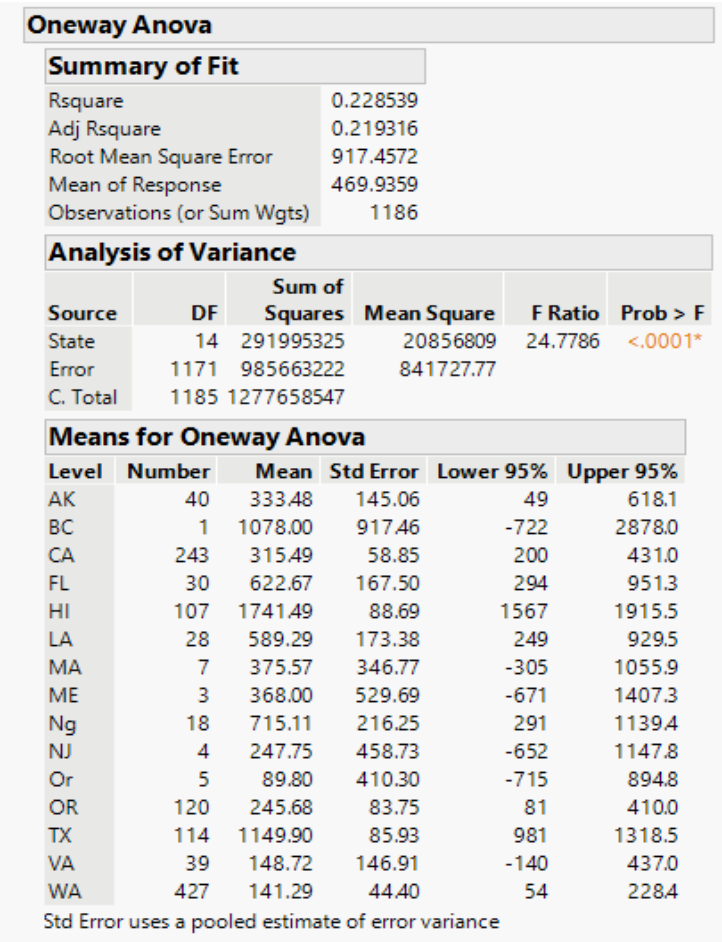
Figure 3.4: ANOVA for State vs Total_Debris



First, we focused on the "Analysis of Variance." Looking at the F-Statistic (F-Ratio) in Figure 3.5, this value is 24.7786. This means that our data is significant and allows us to perform different analyses regarding the debris totals against the states.

The p-value (Prob>F) is <0.001 shows that there is something different and not all the means of the marine debris are the same. This is significant according to the p-value. From there, we conducted an "Unequal Variances" test to see if each group's variances were equal using a Levene Test, which did not result in the same as the null hypothesis. The results from the Levene Test, depicted in Figure 3.6, resulted in a p-value of 0.001, which indicates that the variables are unequal, and the Welch Test should be performed. The Welch Test (Figure 3.6) resulted in a significant p-value less than 0.005 (<0.001), so a multiple comparisons test was performed. We can see that Hawaii (HI) is the state with the highest mean value, which means Hawaii has the most debris pollution compared to the rest of the states. Based on this analysis, we would recommend that our clients focus on Hawaii when targeting ocean cleanups for debris. This can be due to the high number of tourists visiting Hawaii which causes an increase in pollution in the local sea water. This could be one of the reasons that cause increased debris pollution in the waters of Hawaii. We can see that the model explains the 22.85% of variation in our dataset. Moreover, we have a large value for F ratio which suggests that our model gives us valid information

Figure 3.5: RSquare and Mean values of the ANOVA analysis



Oneway Anova

Summary of Fit

Rsquare	0.228539
Adj Rsquare	0.219316
Root Mean Square Error	917.4572
Mean of Response	469.9359
Observations (or Sum Wgts)	1186

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
State	14	291995325	20856809	24.7786	<.0001*
Error	1171	985663222	841727.77		
C. Total	1185	1277658547			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
AK	40	333.48	145.06	49	618.1
BC	1	1078.00	917.46	-722	2878.0
CA	243	315.49	58.85	200	431.0
FL	30	622.67	167.50	294	951.3
HI	107	1741.49	88.69	1567	1915.5
LA	28	589.29	173.38	249	929.5
MA	7	375.57	346.77	-305	1055.9
ME	3	368.00	529.69	-671	1407.3
Ng	18	715.11	216.25	291	1139.4
NJ	4	247.75	458.73	-652	1147.8
Or	5	89.80	410.30	-715	894.8
OR	120	245.68	83.75	81	410.0
TX	114	1149.90	85.93	981	1318.5
VA	39	148.72	146.91	-140	437.0
WA	427	141.29	44.40	54	228.4

Std Error uses a pooled estimate of error variance

Figure 3.6: Welch's Test, Levene Test, Bartlett Test

Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	9.7666	13	1171	<.0001*
Brown-Forsythe	20.8004	13	1171	<.0001*
Levene	45.6237	13	1171	<.0001*
Bartlett	116.6704	13	.	<.0001*

Warning: Small sample sizes. Use Caution.

Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
19.8570	13	46.163	<.0001*

In Figure 3.7, we have 3 different groups: A, B and C. As a result, HI, BC, ME and NJ are in the same group, A, which means that they are not significantly different. Moving onto multiple comparisons, we generated an "All Pairs, Tukey HSD" test, and looked at the "Means Comparisons," focusing on the "Connecting Letters Report" generated by ANOVA analysis. We can conclude that two states, Hawaii, and Texas, are significantly different in terms of the total debris. It generated a few groups based on the mean locations alongside the levels (total debris). It tells us that some of the groups are overlapping, and they are not significantly different from each other

Figure 3.7: Connection Letters Report

Connecting Letters Report

Level				Mean
HI	A			1741.4860
TX		B		1149.9035
BC	A	B	C	1078.0000
Ng		B	C	715.1111
FL		B	C	622.6667
LA		B	C	589.2857
MA		B	C	375.5714
ME	A	B	C	368.0000
AK			C	333.4750
CA			C	315.4856
NJ	A	B	C	247.7500
OR			C	245.6833
VA			C	148.7179
WA			C	141.2857
Or		B	C	89.8000

Levels not connected by same letter are significantly different.

4. Interdependence Analysis

This section

4.1 Principal Component Analysis

Next, we ran a Principal Component analysis (PCA) on the correlations, the results of which can be observed in Figure 4.1. When looking at the Eigenvalues, we focused on values >=1. Looking at the results in Figure 4.1, four values are >=1 (3.4022, 2.2118, 1.0456, 1.0370), so four principal components should be used. Looking at the Eigenvalues report, we can also observe cumulative percentages of variation. Values that have a cumulative percentage <70 are deemed significant. In Figure 4.2 under the cum percent column, four values are <70, again indicating the number of components to use should be four. This can also be proved again when analyzing the results of the scree plot, shown in Figure 4.3, which shows an “elbow” at four components. This is significant because we can inform our client to focus on this portion of the data. We focused on the top 2 principal components in order to inform the client about which items to pinpoint for ocean cleanups. We observed

cluster points across the types of debris shown in the loading plot in Figure 4.1. We identified two major clusters within our independent variables. The first cluster included Soft_Plastics_Total, Food_Items_Total, Hard_Plastics_Total, Bottles_Total, Misc_Hard_Goods_Total, and Misc_Soft_Goods_Total. The other cluster centered around the flux datasets, which identify subgroups of micro particles within the debris sets and included Soft_Plastics_Total_Flux, Food_Items_Total_Flux, Hard_Plastics_Total_Flux, Bottles_Total_Flux, Misc_Hard_Goods_Total_Flux, and Misc_Soft_Goods_Total_Flux. We expected soft plastics and other total items to be grouped together due to their categorical similarities, as well as the flux categories, since all flux variables are measures of micro particles found

Figure 4.1: PCA Analysis

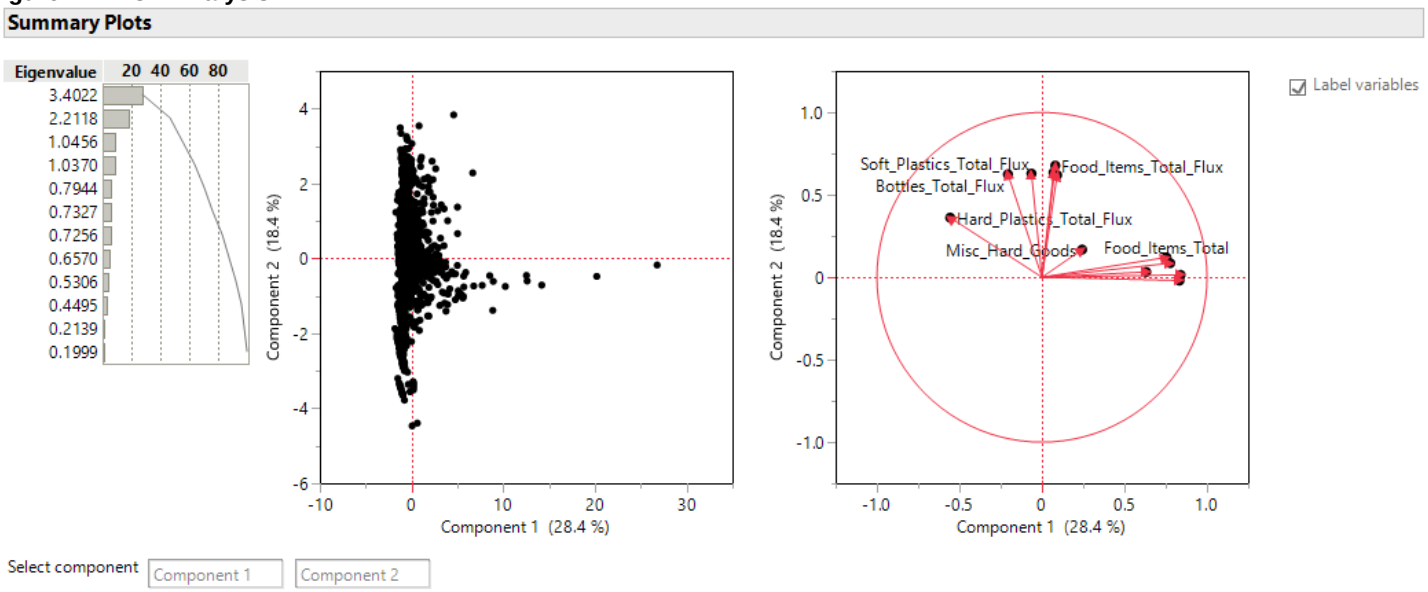
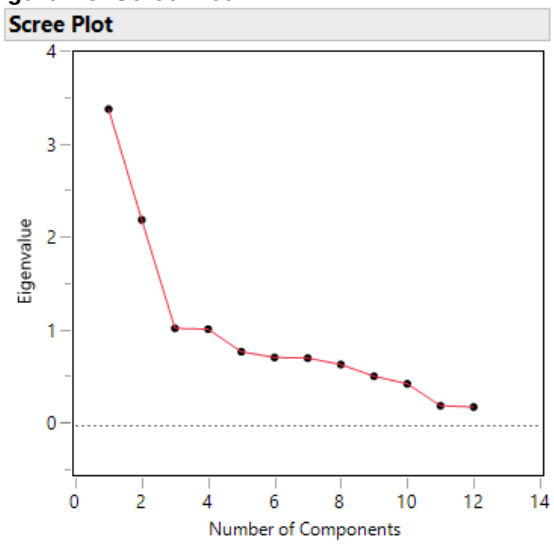


Figure 4.2: Eigenvalues

Eigenvalues

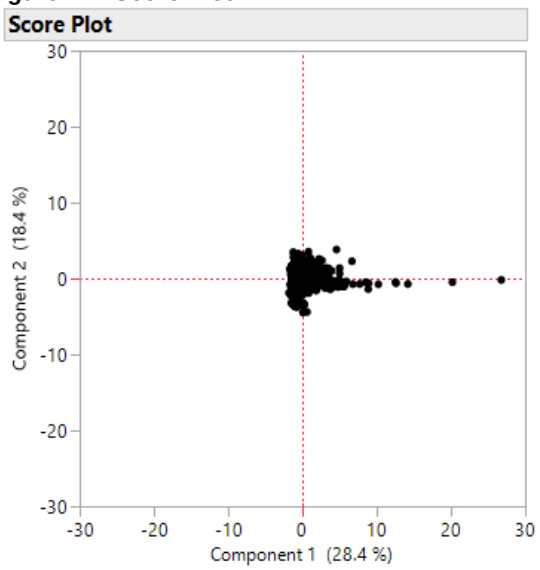
Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	3.4022	28.352		28.352
2	2.2118	18.432		46.783
3	1.0456	8.713		55.496
4	1.0370	8.641		64.138
5	0.7944	6.620		70.757
6	0.7327	6.106		76.864
7	0.7256	6.047		82.910
8	0.6570	5.475		88.385
9	0.5306	4.421		92.807
10	0.4495	3.746		96.552
11	0.2139	1.782		98.334
12	0.1999	1.666		100.000

Figure 4.3: Scree Plot



We found two significant outliers within the components found in the score plot in Figure 4.4: Rows 103 and 89. When analyzing the reasons why these outliers exist, we combed through the original data and found both rows have locations in Maui. The total debris found for both rows were 12410 and 8622, respectively, and presented the highest total debris numbers found within all rows of the dataset. This is good information because all of these outliers are locations in Hawaii. This pinpoints areas of high debris for our client to focus their efforts on and make the most of their ocean clean resources (in terms of prioritization).

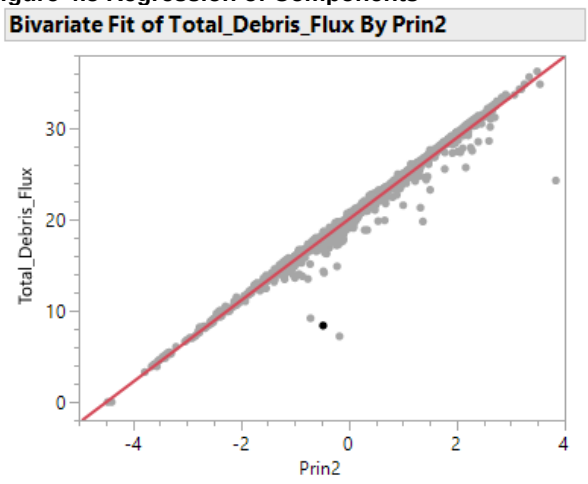
Figure 4.4: Score Plot

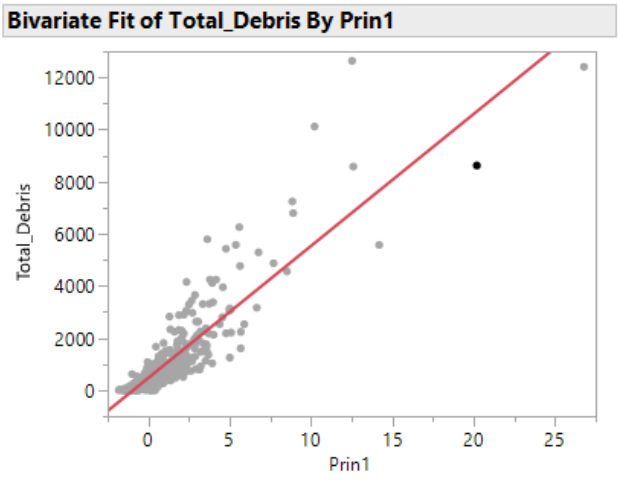


As seen in Figure 4.5, running regressions on the two main components compared with the total debris and total flux values yielded significant RSquare values of 0.809147 and 0.982302, respectively. Comparing the RSquare values between the original and the two main components. We also looked at the impact of

Response_Total Debris to flux items. This allowed the team to focus on the impact of flux items (micro debris) to its overall impact of micro marine debris and its relationship to the overall Total_Debris category.

Figure 4.5 Regression of Components





Linear Fit

Linear Fit

Total_Debris_Flux = 20.034654 + 4.4642387*Prin2

Summary of Fit

RSquare	0.982302
RSquare Adj	0.982287
Root Mean Square Error	0.891541
Mean of Response	20.03465
Observations (or Sum Wgts)	1186

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	52234.557	52234.6	65716.66
Error	1184	941.096	0.794845	Prob > F
C. Total	1185	53175.654		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	20.034654	0.025888	773.90	<.0001*
Prin2	4.4642387	0.017414	256.35	<.0001*

Linear Fit

Linear Fit

Total_Debris = 469.93592 + 506.38767*Prin1

Summary of Fit

RSquare	0.809147
RSquare Adj	0.808986
Root Mean Square Error	453.8174
Mean of Response	469.9359
Observations (or Sum Wgts)	1186

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1033813436	1.0338e+9	5019.724
Error	1184	243845111	205950.26	Prob > F
C. Total	1185	1277658547		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	469.93592	13.17768	35.66	<.0001*
Prin1	506.38767	7.14732	70.85	<.0001*

4.2 Hierarchical Clustering

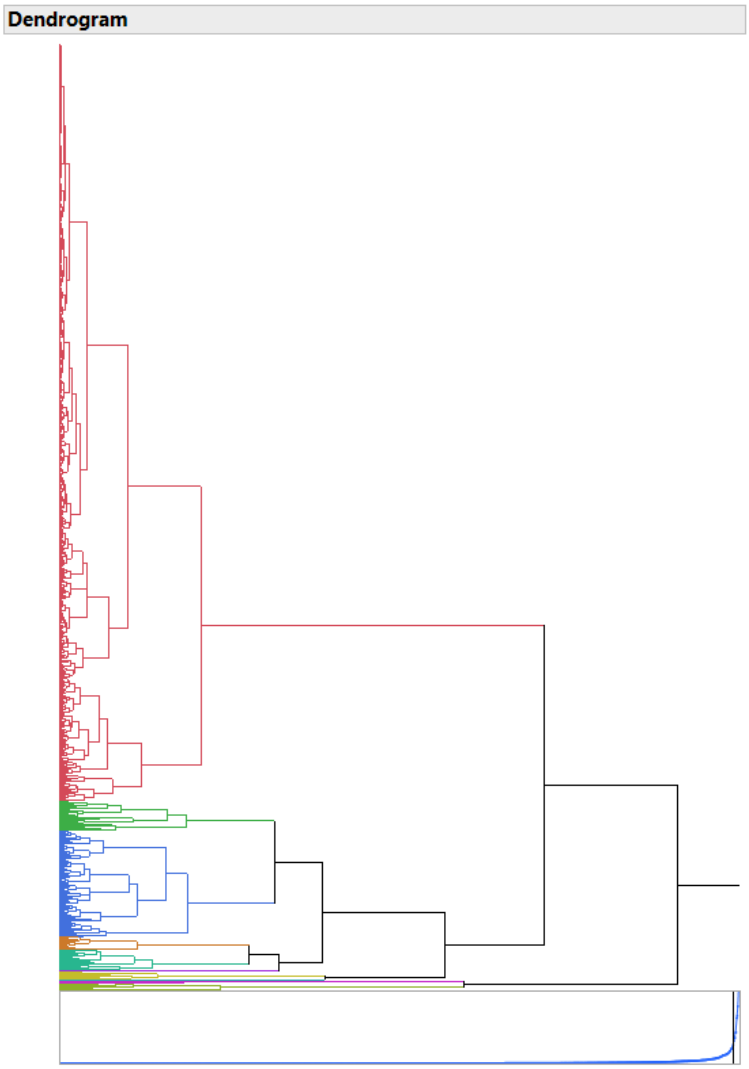
We performed hierarchical clustering, which enabled us to analyze clusters of types of total marine debris specific to the state in which the marine debris was located. We performed a hierarchical clustering method in Figure 4.6. When analyzing the hierarchical cluster we observed the number of debris item types to the states, more specifically on the shorelines of each state. First off, after plotting the dendrogram using the Ward method, we created a scree plot. This gave valuable information as it looked to me as if the elbow was right about 10 (natural break), showing that 10 clusters would be a good choice for this analysis.

This scree plot helped us strategically cluster the types of marine debris totals in an organized manner to assist our client with mapping out states to focus on for ocean clean ups. The various colors indicate the 10 clusters of debris types by state.

Another interesting observation from the dendrogram, is how the color helps to illustrate the number of points (visually) within each cluster. This will be better displayed within the constellation plot below, but we can observe where most marine debris fall within the 10 clusters overall. This gives our client an idea of the marine debris at higher volumes within specific locations overall.

The constellation plot within Figure 4.6 shows the 10 clusters more clearly. For example, the red clusters represent total hard plastics (Hard_Plastics_Total) found and in what prevalence by state, indicated by the labels. This is useful information for our company because specific debris items can be found in different or the same states.

Figure 4.6: Hierarchical Clustering



4.3 Parallel Plot and Scatterplot Matrix

Next, we observed the parallel plot of the hierarchical clustered data. Through this analysis we observed the various marine debris items and how they peak or dip within their various clusters. The curves of each parallel plot (10 total plots) correspond to the various ocean clean up events and their prospective marine debris item can be seen in Figure 4.7. In the first parallel plot (red), we observed that there is a good mix of each type of marine debris focused on within our analysis, with small peaks at HardPlastics and Bottles. Cluster 2 shows a peak at FoodItems and a smaller peak at HardPlastics. This could show that there is a good amount of these items within most of the clustered ocean clean up spots. Cluster 3 was very similar but peaked at SoftPlastics. Cluster 4 peaks at Bottles significantly. Hence, highlighting a potential need for our client to pinpoint these areas as places where bottles are frequently disposed of and target specific locations (based on locations included in the cluster). Cluster 5 has a peak at Bottles too, but also a significant peak at HardPlastics. The HardPlastics came up the most within this dataset and should be noted for collection targeting as well.

Singletons occurred in plots 6 and 8. This is significant because each plot peaked at a different location. Noting a place that could be significant to target for marine debris cleanups. Bottles was a peak in plot 6 and Misc

Hard Goods came up in plot 8. Bottles are consistently plotted within other clusters and we will advise our client to focus within these areas as well for marine debris point-source material targeting.

Plots 7, 9, and 10 are unique. Each has an okay amount of cluster data, yet the peaks are more random than the other plots. Still, we can notice that Hard Plastics and Bottles have high totals in these plots, again pointing to an emphasis of these items for ocean cleanups.

Figure 4.8 displays the scatterplot matrix for the dataset, which helps to determine levels of relationships between variables (marine debris types and their relationship with one another at ocean cleanups). Specifically, here, we are placing an emphasis on pairs of variables and clusters plotted across the two variables.

Figure 4.8 consistently shows Hard_Plastics_Total to show up in a strong/tight cluster across each item that it is paired with across the marine debris items. This makes sense as our original dataset showed high volumes of Hard_Plastics_Total. Misc_Soft_Goods consistently has outliers with all the other variables, signaling that this category could be less frequent in terms of clusters, but still frequent enough to consistently show up across the variable groups. We will instruct the client to rethink where in terms these outliers could be sorted into categories and if there are specific reasons why these types consistently appear. It could be that these are larger debris items that occur less frequently, yet still appear regardless. Bottles and FoodItems also have recurring clusters across the Scatterplot Matrix, potentially indicating their association with other marine debris items. For example, a Bottle might have had a drink originally inside of it and it could have been paired with a FoodItem that was packaged in food packaging. This “meal” may have contributed to both categories and could point to a strong relationship across these categories.

Figure 4.7: Parallel Coordinates Plot

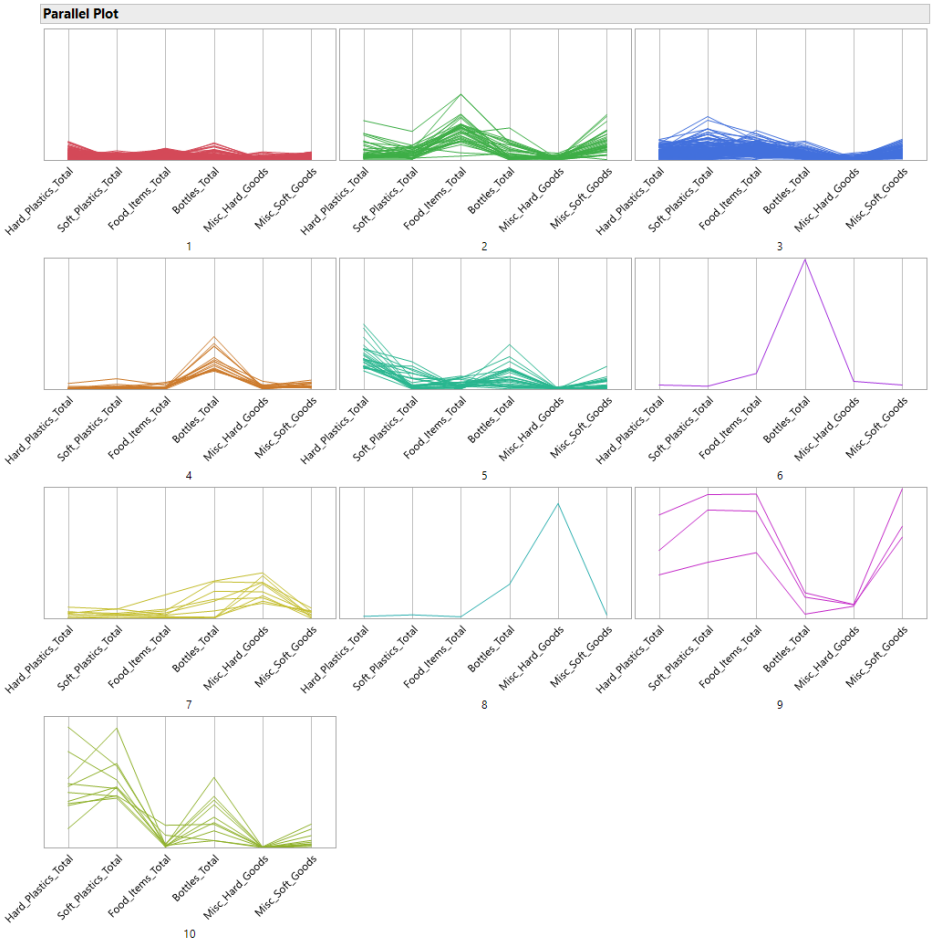
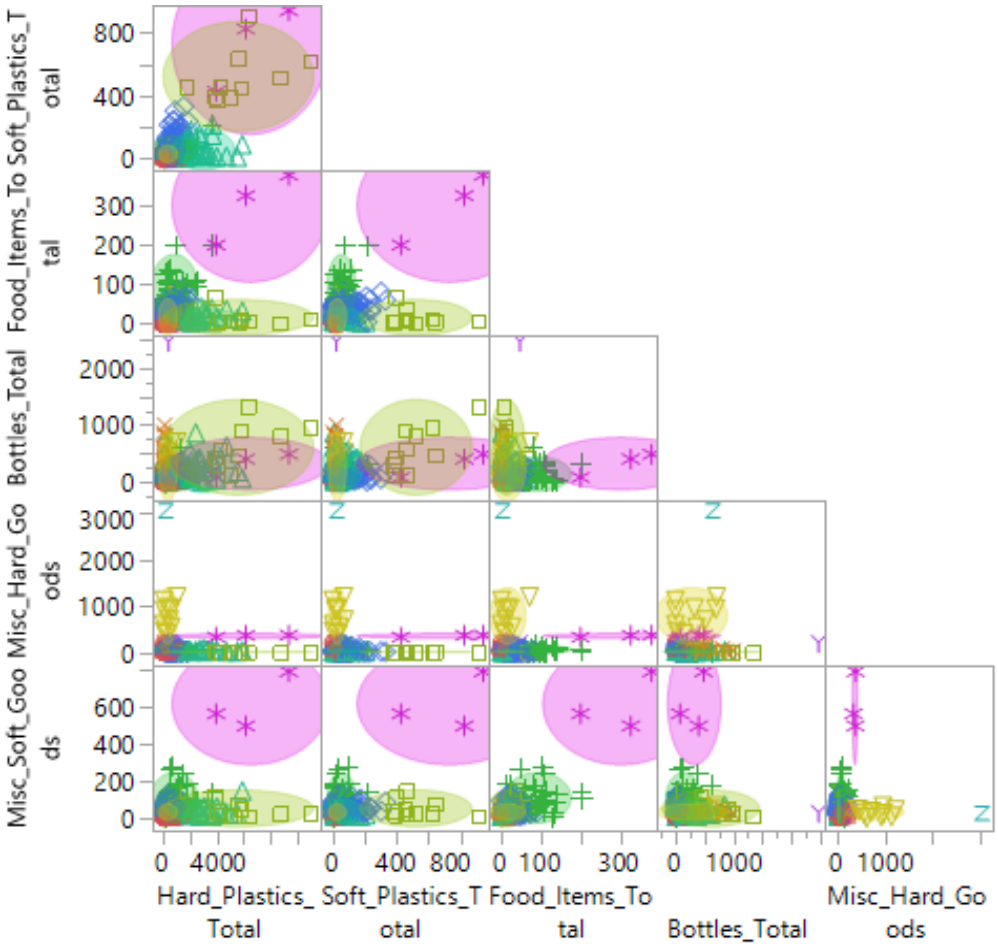


Figure 4.8: Scatterplot Matrix



4.4 ANOVA for Hierarchical Clustering analysis

ANOVA testing for the dataset's continuous variables is displayed in Figure 4.9. Total Debris clusters across marine debris types can be observed. Clusters illustrate where their means are situated in comparison to the other clusters. These clusters are like the clusters plotted above in Figures 4.7 & 4.8. Both show significant/tight clustering within clusters 2 & 3, with no significant outliers in the data (potentially a well-rounded mean for most of the items across this cluster). This aligns well with the RSquare value given in the summary of fit, which focuses on the proportion of variance for the ANOVA for Hierarchical Clustering. The RSquare is 0.786764. This is a significant RSquare value and points to significance in regard to the totals of specific types of marine plastic that make up the total debris for each cluster instance. In terms of the connecting letters report, Figure 4.10 illustrates how each of the 10 clusters are related to each other. There was no overlap between clusters. For example, A & B did not show any connections between each other. This is significant because each marine debris item is independent of one another, and the varying means shown in

Figure 4.10 influence each cluster uniquely. Our client should note these differences when focusing on clusters, as each instance is significantly different.

Figure 4.9: ANOVA for Hierarchical Clustering

Summary of Fit

Rsquare	0.786764
Adj Rsquare	0.785132
Root Mean Square Error	481.32
Mean of Response	469.9359
Observations (or Sum Wgts)	1186

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Cluster	9	1005215891	111690655	482.1132	<.0001*
Error	1176	272442657	23166.93		
C. Total	1185	1277658547			

Figure 4.10: Connecting Letters Report

Connecting Letters Report

Level							Mean
9	A						8868.6667
10		B					6840.5000
8			C				3952.0000
5			C				3378.0769
6			C				3166.0000
7				D			1489.3000
2				D			1273.5135
3					E		949.4812
4					E		771.3750
1						F	175.7102

Levels not connected by same letter are significantly different.

4.5 K-Means Clustering Analysis

Through computing K-Means clustering on our marine data, we uncovered valuable insights related to marine debris clustering that can be compared to our hierarchical clustering analysis. To perform the k-means clustering, we used the same independent variables that were used for hierarchical clustering. The results of our clusters and the different means between them are listed in Figure 4.11. When prompted to determine the range of clusters to be considered, we specified a range of 3 to 10 clusters, as 10 was the optimal number of clusters to use according to the scree plot in Figure 4.6. Again, the optimal number of clusters (Optimal CCC) to use in k-means clustering was 10, as shown in Figure 4.12. These optimal 10 clusters will allow us to strategically guide our client to focus on specific marine debris items. More specifically, which clusters have peak amounts of specific items and what high volume marine debris items to target.

Figure 4.11: Cluster Summary

Cluster Summary

Cluster	Count	Step	Criterion
1	992	7	0
2	1		
3	1		
4	1		
5	1		
6	8		
7	42		
8	9		
9	130		
10	1		

Cluster Means

Cluster	Hard_Plastics_Total	Soft_Plastics_Total	Food_Items_Total	Bottles_Total	Misc_Hard_Goods	Misc_Soft_Goods
1	142.402218	10.1703629	5.60685484	15.6834677	11.4798387	7.59677419
2	9442	942	378	488	371	789
3	390	23	48	2468	211	26
4	6206	825	326	404	366	495
5	3979	427	200	82	325	561
6	396.625	31.375	20	327.875	880.25	24.375
7	1600.52381	41.547619	10.8095238	405.02381	31.7380952	28.1904762
8	6069.44444	528.333333	13.2222222	664	6.22222222	38
9	906.123077	81.0461538	54.4923077	111.646154	38.9692308	68.9384615
10	188	27	5	649	3062	21

Figure 4.12: Cluster Comparison

Cluster Comparison

Method	NCluster	CCC	Best
K Means Cluster	3	-22.054	
K Means Cluster	4	-9.202	
K Means Cluster	5	-10.047	
K Means Cluster	6	-15.427	
K Means Cluster	7	-8.8914	
K Means Cluster	8	11.7292	
K Means Cluster	9	17.4078	
K Means Cluster	10	17.7682	Optimal CCC

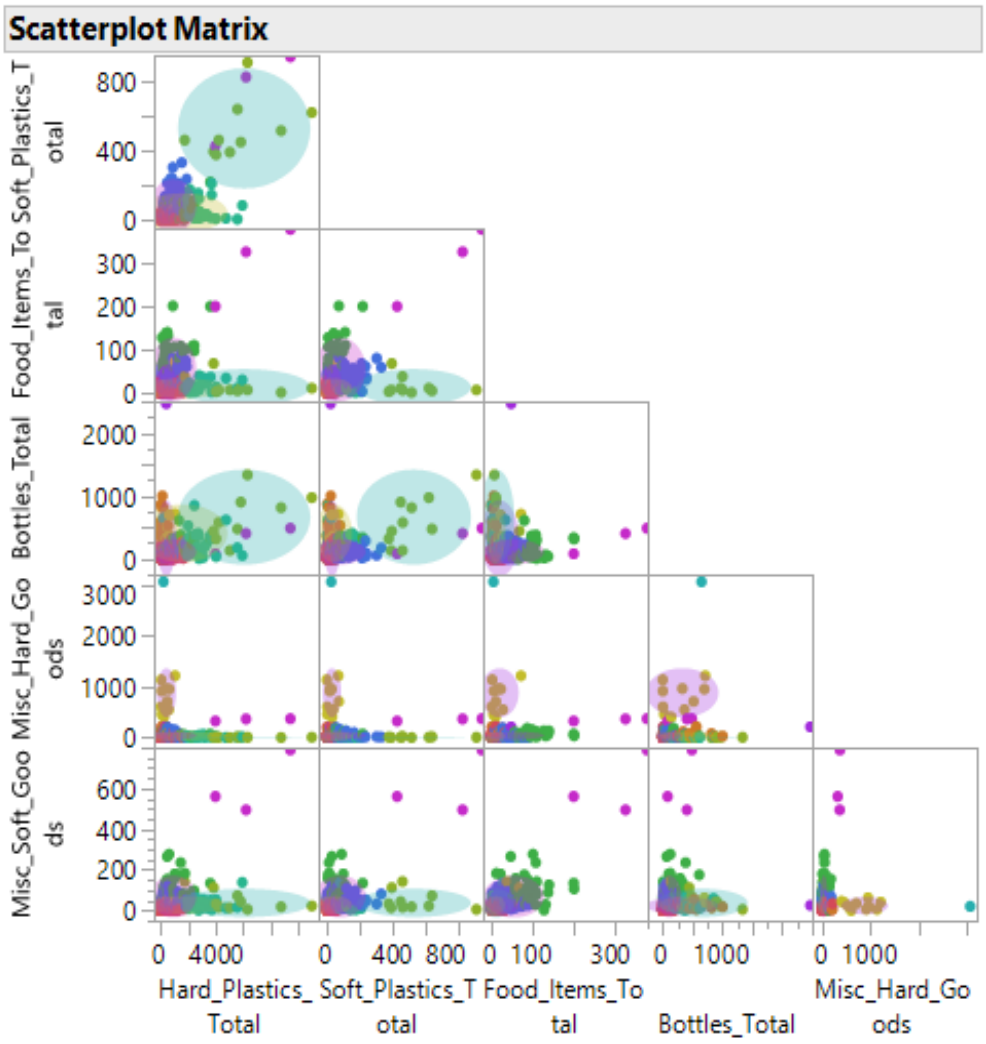
4.6 K-Means Scatterplot Matrix

Next, we focused on the scatterplot matrix and compared it to the scatterplot matrix of the hierarchical cluster. The K-Means Scatterplot Matrix is plotted in Figure 4.13. It's interesting that the scatterplot for the k-means did not seem to be as spread out as the hierarchical cluster scatterplot matrix.

The k-means scatter plot seems to cluster the marine debris more compactly, which from an observational standpoint, could help our client create higher-level marine debris to target more marine debris that are commonly clustered together. Again, the HardPlastics seems defined even more in the Figure 4.13 K-Means Scatterplot Matrix, as opposed to the Hierarchical Clustering Scatterplot Matrix in Figure 4.8, possibly pointing to an opportunity for our client to focus their efforts highly on HardPlastic collection and look for regional recyclers to send the HardPlastic items.

The Miscellaneous related marine debris categories were still observed as outliers in Figure 4.13. This is good information because our client can use this mindset to identify which points are the outliers/singletons. From there, the wholesale distributor can strategically pair the singletons categories such as HardPlastics that are specific to a location. This will allow our client to effectively use their resources when collecting the marine debris and group in the outliers for efficiency purposes.

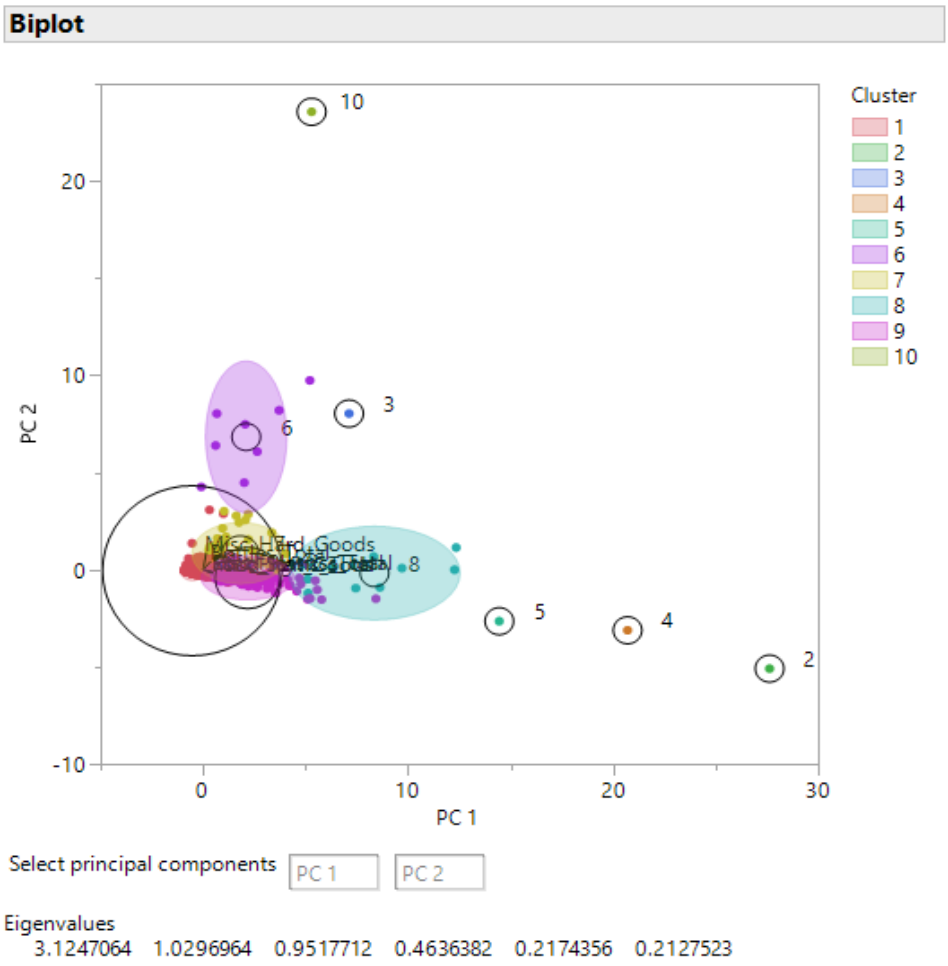
Figure 4.13: K-Means Scatterplot Matrix



4.7 K-Means Biplot

Last, we focused on the K-Means Bilpot in Figure 4.14, to gain insights into how the points are distributed with the first two principal components and see the clusters of the first two principal components. Focusing on the shaded regions (95% confidence interval areas), it can be observed that there is a lot of overlap with clusters 1,7, and 9. This could further show our client which clustered locations of marine to group together for ocean cleanup activities. These clusters do not have any outliers but signal high levels of marine debris pollution across all of our marine debris categories. The singleton has smaller circles with their small sample size. This is another tool for our clients to correctly identify the clusters of marine debris that don't have as much traction with certain debris pollution.

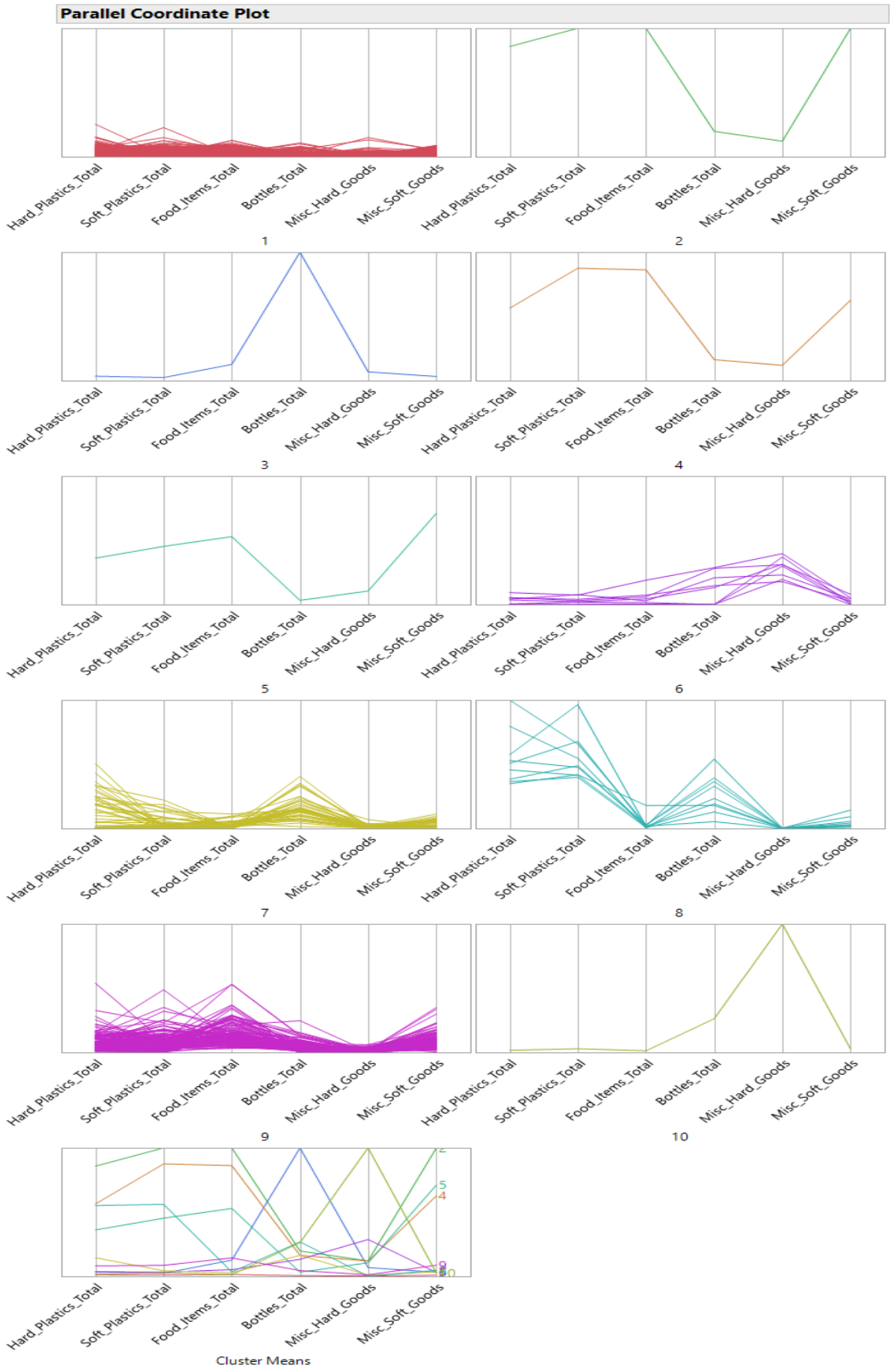
Figure 4.14: K-Means Biplot



4.8: K-Means Parallel Plot

Cluster 1 parallel plot is like the one of hierarchical clustering. We noticed that these locations have small values of debris pollution for all categories. Cluster 2 we can see that we have a singleton which represents row 103 in our data. Very high pollution totals of almost all variables can be noted and row 103 can be noted as outlier. Cluster 3 is also a singleton and represents row 1002 in our dataset. The location of cluster 3 has a very high amount of Bottle_Total pollution. Cluster 4 is also a singleton and represents row 89 in our dataset. The location which is represented by cluster 4 has high values of Soft_Plastics_Total and Food_Items_Total. Cluster 5 is also a singleton which represents row 35 in our data dataset. The location which is represented by cluster 5 has some middle-to-high values of Hard_Plastics_Total, Soft_Plastics_Total, Food_Items_Total and Misc_Soft_Goods. Parallel plot of cluster 6 pinpoints that these specific locations tend to have higher amounts of Misc_Hard_Goods pollution. On cluster 7 that Bottles_Total and Hard_Plastics debris is more common to find. On cluster 8 there are some high values of Soft_Plastics and Bottles_Total pollution. Cluster 9 represents locations where pollution of Food_Items are noticed with the highest frequencies. Cluster 10 is also a singleton and represents row 280 in our dataset. It can tell our client that the specific location has the highest amount of Misc_Hard_Goods pollution in our dataset. Parallel plots can help us as a company to advise our clients and give more information on which pollution collecting methods are going to be more efficient since we can track which type of pollution is more frequent in those specific locations.

Figure 4.15: K-Means Parallel Plots



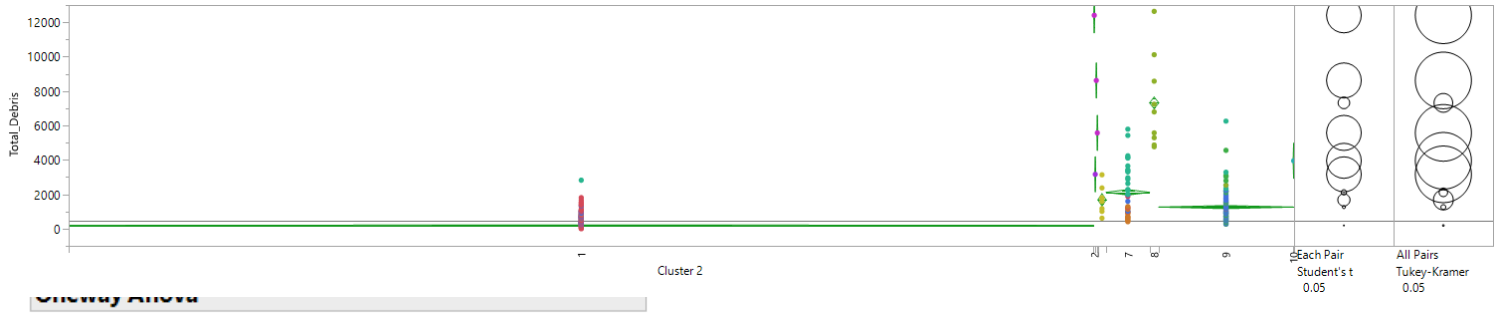
4.9 ANOVA for K-Means Clustering

ANOVA or K-Means Clustering for the dataset's continuous variables is displayed in Figure 4.16. Total Debris clusters across marine debris types can be observed with clustering. These clusters are similar to the clusters plotted above in Figures 4.13 & 4.15. Significant/tight clustering within are shown within 1,7, & 9, with no significant outliers in these clusters. This aligns with the same results we gathered in analyzing K-means parallel plots. The singletons such as cluster 2, show the highest mean values. Even though this represents one set of locations, this is significant due to the high presence of marine debris pollution.

Again, this aligns well with the RSquare value given in the summary of fit, which focuses on the proportion of variance for the ANOVA for Hierarchical Clustering. The RSquare is 0.751349. This is a significant RSquare value and points to significance in regard to the totals of specific types of marine plastics.

In terms of the connecting letters report, Figure 4.17 again illustrates how each of the 10 clusters are related to each other. Unlike the hierarchical clustering, there was some overlapping with the connecting letters clustering. For example, Cluster 10 had overlap with C & D and Cluster 3 had overlap with D & E. This is significant because we can suggest common strategies to collect marine debris within these overlapping clusters. Still, there are a fair amount of non-connecting letters that point to non-significant correlations across clusters. For example, cluster 1 and cluster 2 are significantly different from the other clusters because they only have one connecting letter. Hence, showing significantly different groupings of marine debris within these clusters. Our client should note these similarities and differences when focusing on clusters, as each instance is significantly different.

Figure 4.16: ANOVA analysis for K-Means Clustering



Summary of Fit

Rsquare	0.751349
Adj Rsquare	0.749446
Root Mean Square Error	519.7552
Mean of Response	469.9359
Observations (or Sum Wgts)	1186

Analysis of Variance

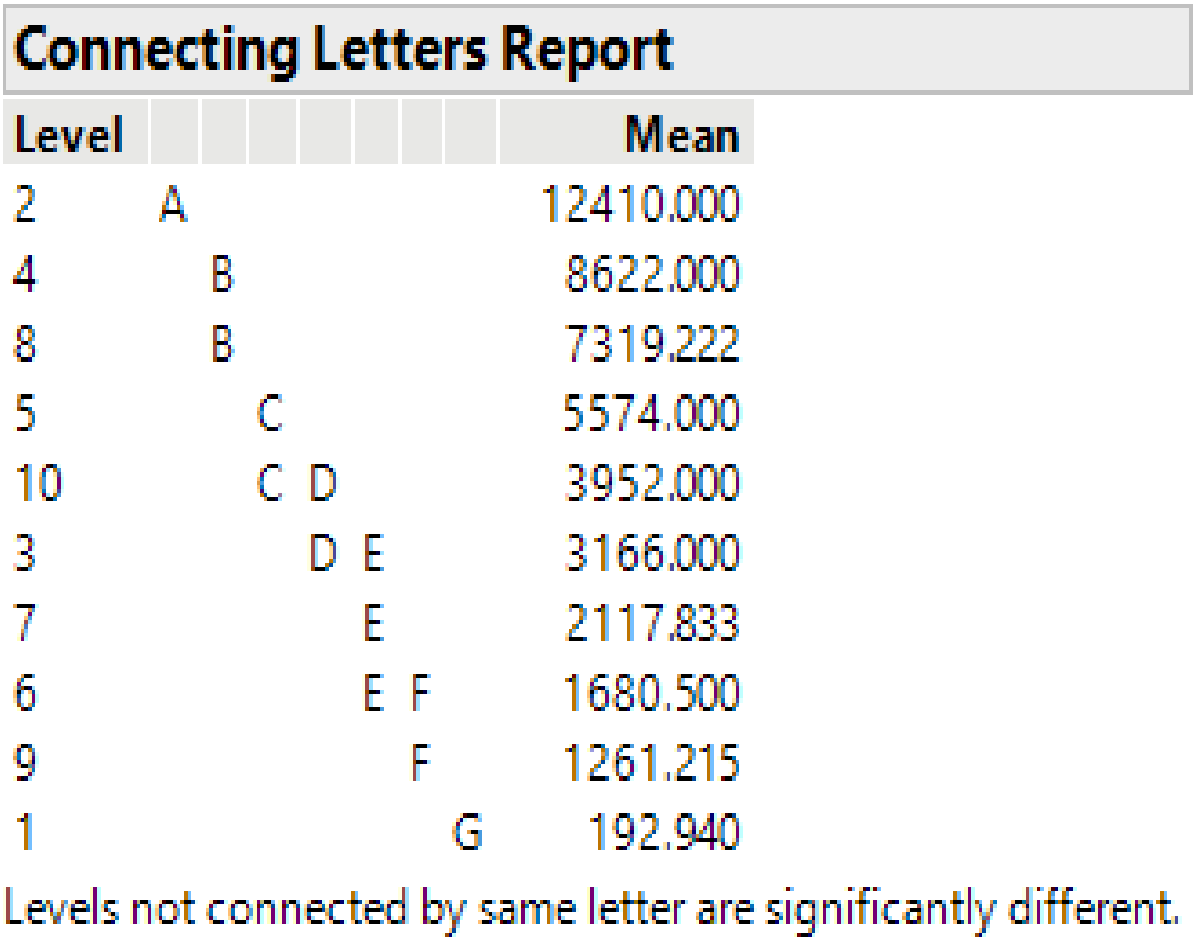
Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Cluster 2	9	959967483	106663054	394.8356	<.0001*
Error	1176	317691064	27014546		
C. Total	1185	1277658547			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
1	992	192.9	16.50	161	225
2	1	12410.0	519.76	11390	13430
3	1	3166.0	519.76	2146	4186
4	1	8622.0	519.76	7602	9642
5	1	5574.0	519.76	4554	6594
6	8	1680.5	183.76	1320	2041
7	42	2117.8	80.20	1960	2275
8	9	7319.2	173.25	6979	7659
9	130	1261.2	45.59	1172	1351
10	1	3952.0	519.76	2932	4972

Std Error uses a pooled estimate of error variance

Figure 4.17: K-Means Connecting Letters



Connecting Letters Report

Level								Mean
2	A							12410.000
4		B						8622.000
8		B						7319.222
5			C					5574.000
10			C	D				3952.000
3				D	E			3166.000
7					E			2117.833
6					E	F		1680.500
9						F		1261.215
1							G	192.940

Levels not connected by same letter are significantly different.

5. Model Comparison

The purpose of this model comparison is to compare the marine ocean variables and see which types of predictive modeling will be the best for determining factors such as type of marine debris for our client to target. The model comparison will allow for effective usage of variables and link the significance of these variables to

optimized modeling techniques. For example, model comparison allows the user to compare two models where comparing the RSquare values would be obsolete. In our case, a binary value and a continuous value.

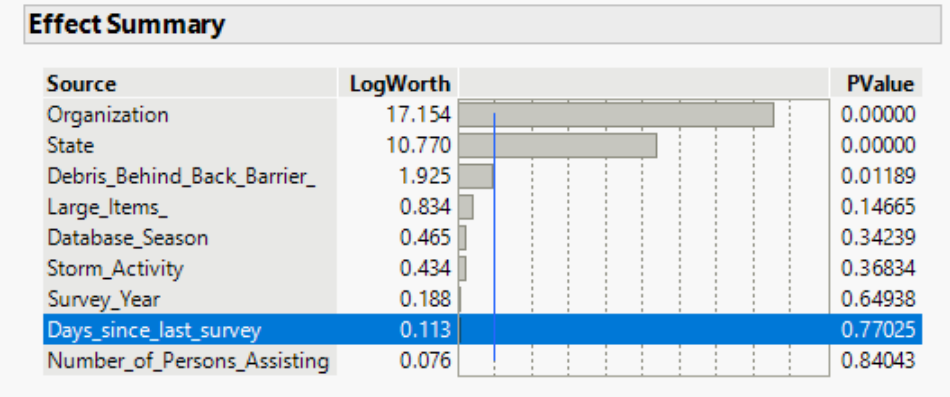
To begin our model comparison, Y Binary and Validation columns must first be created. Using Total_Debris as our dependent variable Y, we were able to use the average value of the column (using Analyze > Distribution), 469.93592, to calculate our Y Binary column. A conditional formula was used to output rows stating whether values were “High” or “Low” in relation to the Total_Debris column average. The Validation column was randomly generated by inputting 0.667 into the Training and 0.333 into the Validation sets. Models to be used include: Linear regression, decision tree, neural networks, and support vector machines. For each model, ROC and Lift curves will be generated to compare between models, with the AUC (Area Under the Curve) being used as a performance measure. Since the ROC has probabilistic interpretation properties, the higher the AUC, the better the model prediction.

5.1 Predictive Models

5.1.1 Logistic Regression

The logistic regression gives us the relationship between the Y Binary (High total debris/Low total debris amounts) with the independent variables such as Organization, State, Debris_Behind_Back_Barrier_, Large_Items_, Database_Season, Storm_Activity, Survey_Year, Days_Since_Last_Survey, and Number_of_Persons_Assisting. We can see that Organization, State, and Debris_Behind_Back_Barrier_ have a significant relationship with High or Low amounts of Total Debris since they are the variables which have LogWorth value higher than 1 which can be seen on Figure 5.1. Therefore, we could conclude that the effectiveness of Organizations and their locations (States) could result in the gathering of higher total debris amounts. Moreover, we can see that there is a high correlation in geographical locations. All locations studied included beaches that have a backbarrier, either natural or man-made; a diagram of which can be found in Figure 3 within Appendix A (Slatt, 2013). The lack of waves and a high current to the area due to the back-barrier result in higher collections of total debris as an accumulation builds up over time. Additionally, we can see in Figure 5.1 that our logistic regression has a high Rsquare value of 0.6299 which means that our model is able to explain the 62.99% of the data variability.

Figure 5.1: Logistic Regression Results for variables



Effect Summary

Source	LogWorth	PValue
Organization	17.154	0.00000
State	10.770	0.00000
Debris_Behind_Back_Barrier_	1.925	0.01189
Large_Items_	0.834	0.14665
Database_Season	0.465	0.34239
Storm_Activity	0.434	0.36834
Survey_Year	0.188	0.64938
Days_since_last_survey	0.113	0.77025
Number_of_Persons_Assisting	0.076	0.84043

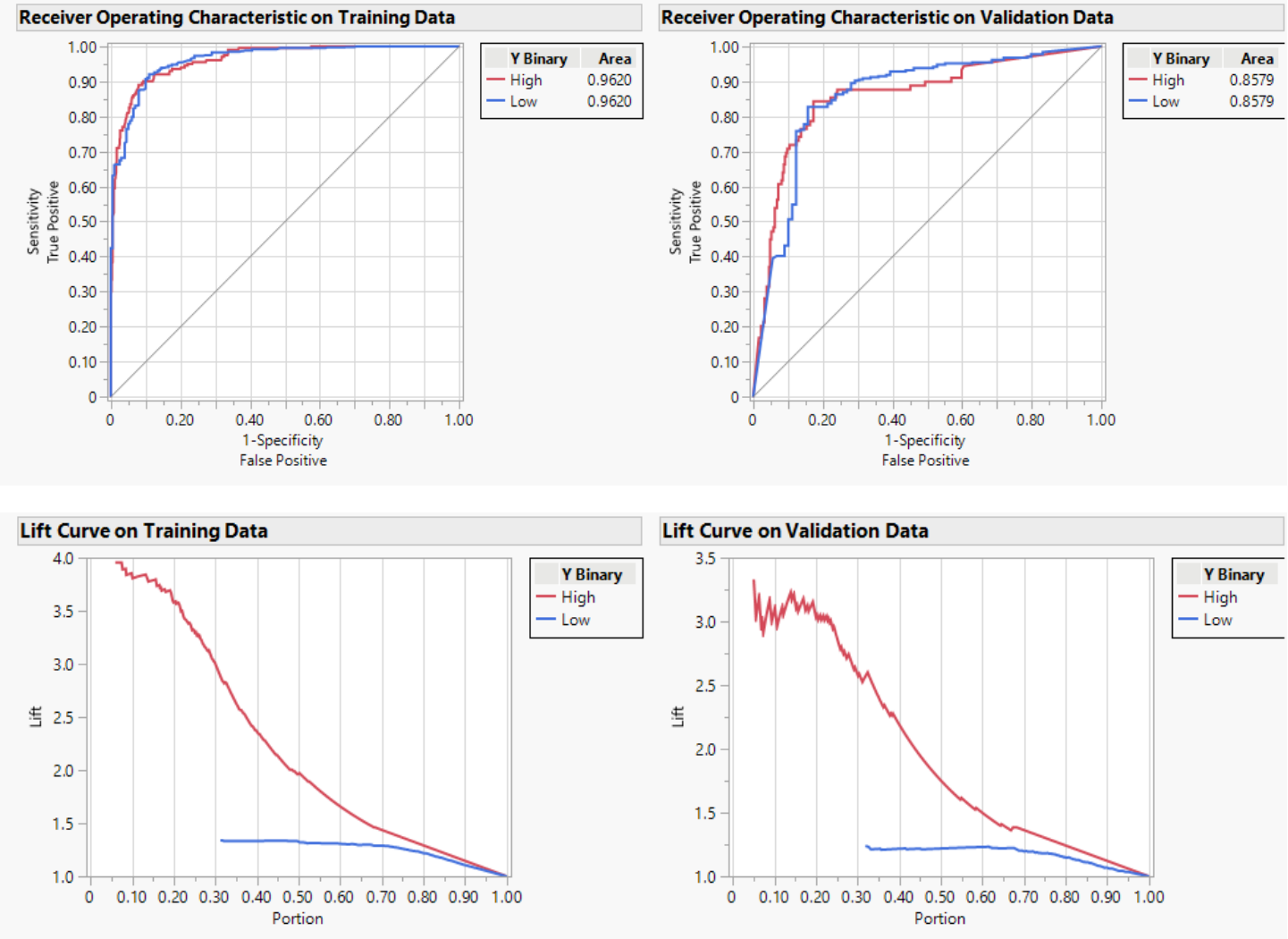
Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	281.72069	157	563.4414	<.0001*
Full	165.54126			
Reduced	447.26195			

RSquare (U)	0.6299
AICc	849.553
BIC	1632.38
Observations (or Sum Wgts)	791

Analyzing the ROC and Lift curves (as seen in Figure 5.2), we can see that for the Training set of the ROC curve the AUC is 0.9620, and for Validation it is 0.8579. In general, the closer the AUC is to 1, the better the model is able to accurately predict true positive values. Training is almost at 1, while Validation is high enough to accurately predict true positives. The curve of the ROC is far away from the middle line, showing a good capacity of the classifier to be chosen in the correct order. Good marine debris classifiers can be justified from this result. The lift curve shows likelihood that variables are contributing to marine debris pollution, based on the portion and lift. A join of the lines at around 0.96 shows the significance (likelihood) of percentage variables that contribute to the marine debris totals (i.e-.Organization, State, Debris_Behind_Back_Barrier_).

Figure 5.2: ROC and Lift Curves for Logistic Regression



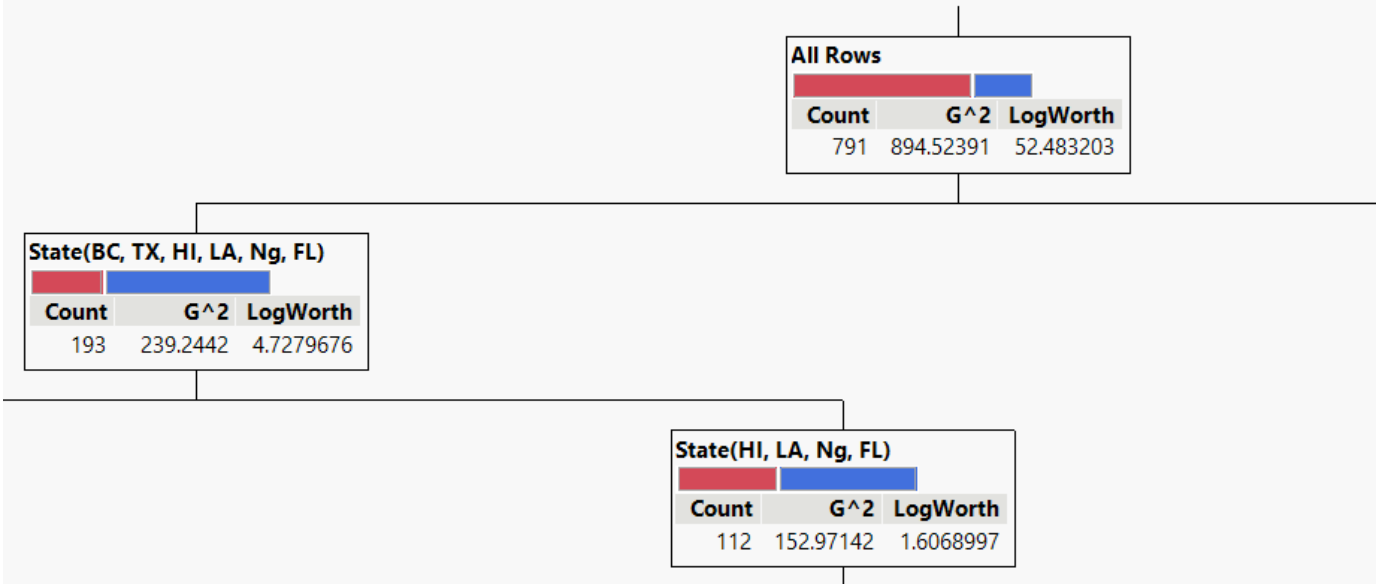
5.1.2 Decision Tree

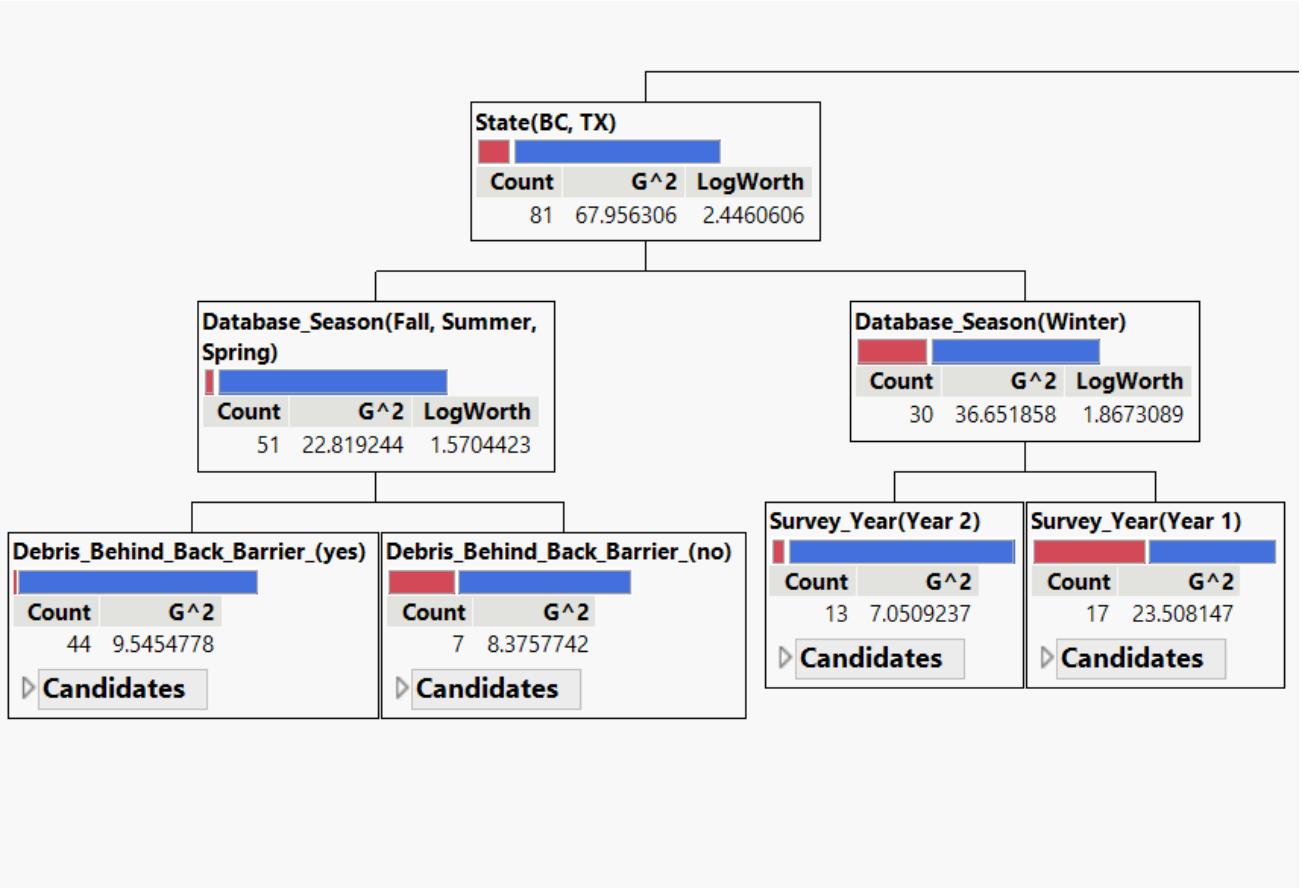
The decision tree analysis allowed us to evaluate options of variables that contribute to the input and output (validation and training) when the data is split according to a parameter. Using Y Binary (threshold) for Y and State, Survey_Year, Season, Large_Items, and Number_Of_Persons_Assisting for X values, we determined that 13 was the optimal number of splits. We determined that 13 was the optimal number of splits because a split of 15 or higher began to split small numbers of people involved in marine debris cleanup (which is shown as insignificant). In addition, the RSquare value became much smaller and pointed to an insignificant split. Splits lower than 13 seemed to leave out information such as types of debris that our client should focus on within each state. Hence pointing to 13 as the optimal split count.

As we began to split the data, we recognized that a significant story about marine debris was being played out and columns of high contributions could be seen. For example, as seen in Figure 5.3, States on the west coast of the United States such as California and Washington had lower “counts” compared to the right side of the split, which included states in the southeastern coastal region of the United States. From this we concluded that marine debris must be compiling heavily in places such as the Gulf of Mexico. In addition, on the left side of the decision tree, winter had a lower count as opposed to the other seasons, with a 50:31 ratio. In British Columbia and Texas, diving further into the split, we can see that 44/51 counts had items behind the back barrier, hence pointing to a more significant flow of marine debris during warmer seasons. The complete version of this decision tree can be viewed in Appendix A as Figure 5.

Figure 5.3: Decision Tree Model for Training and Validation

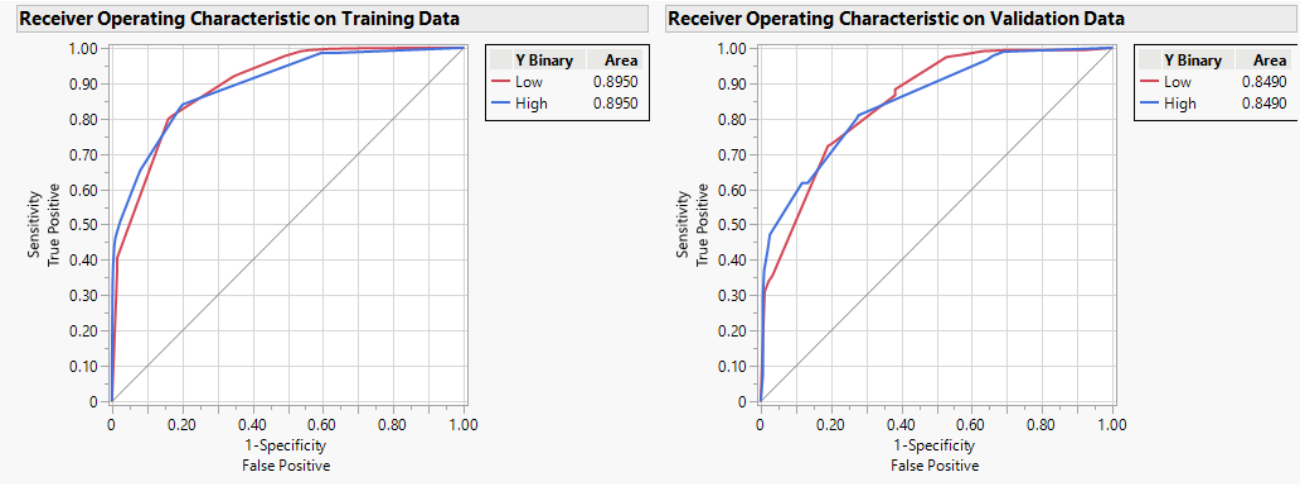
	RSquare	N	Number of Splits
Training	0.439	791	13
Validation	0.312	395	

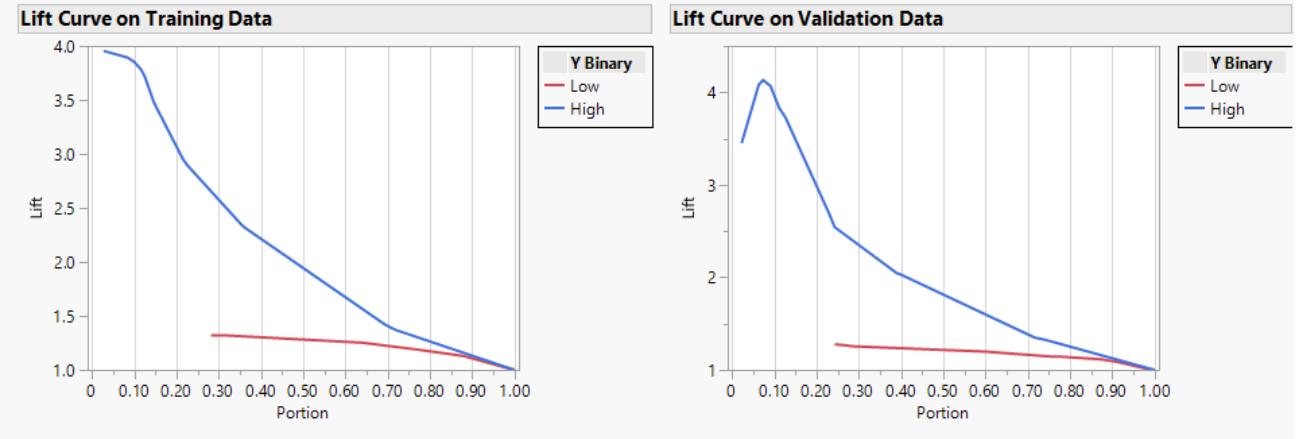




In Figure 5.4, the ROC and Lift curves for the decision tree are listed, with the Training set on the left and Validation on the right. For the ROC curves, the AUC for both Training and Validation are similar, with AUCs of 0.8950 and 0.8490, respectively. The lift curves meet at a point around 0.90. Compared with the linear regression model above, the ROC values are slightly lower, but the values still mean the model is significant when predicting true positive values.

Figure 5.4: ROC and Lift Curves for Decision Tree





Referring to the confusion matrix in Figure 5.5, we can see that there is slight overfitting occurring between the Training and Validation models. The counts for the true negative values are higher by 279 values in the Training model than in the Validation model, and the Training model predicted 97.6% of true negatives, compared to 97.4% of Validation. In the lower right-hand corner of the matrix that depicts the amount of true positive predictions, Training has a count 60 values higher than Validation, and predicted 51% of true positives, compared to 47.2% of Validation.

Figure 5.5: Confusion Matrix for Decision Tree

Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.4388	0.3121	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.5776	0.4318	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3173	0.3670	∑ -Log(ρ[j])/n
RASE	0.3141	0.3365	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.2006	0.2270	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1416	0.1392	∑ (ρ[j]≠ρMax)/n
N	791	395	n

Confusion Matrix

Training

Actual	Predicted Count	
Y Binary	Low	High
Low	577	14
High	98	102

Actual	Predicted Rate	
Y Binary	Low	High
Low	0.976	0.024
High	0.490	0.510

Validation

Actual	Predicted Count	
Y Binary	Low	High
Low	298	8
High	47	42

Actual	Predicted Rate	
Y Binary	Low	High
Low	0.974	0.026
High	0.528	0.472

5.1.3 Neural Network

We ran a boosted neural network analysis in order to improve the performance of our data modeling and to help reduce the overall bias within the data. Here, we used TanH(3)NBoost(24) on Y Binary to output our model. This resulted in a generalized RSquare of 0.7480661 for Training and 0.6101745 for Validation. These results can be found in Figure 5.6. Compared to the other models run, such as the normal model and the weak model, this boosted model had better results when comparing the RSquare values and ROC and lift curves. For the misclassification rates, Training was only 7.46% and Validation was higher, at a rate of 11.39%.

Finally, we ran a neural network boosted analysis Boosting the neural network is a way to improve the performance of the data you are modeling and can help to reduce bias in a dataset. Here, we included 1 TanH neurons and entered 100 for the number of the Models in the Boosting section. This resulted in an RSquare of roughly 0.62 for training and roughly 0.47 for validation. These values are pretty close to the results from the neural networks weak analysis and not far off from the neural networks standard analysis. Hence showing a fair amount of significance across all three of these models. The boosted model had the highest RSquare validation, potentially due to the nature of the boosting method to curb overfitting and usage of bagging. As seen in Figure 5.6, the RASE of the training was 0.24 and the RASE for validation was 0.28. This is a slight change and not very noFigure. The misclassification rate slightly increased from training to validation (0.07 to 0.11). In addition, Figure 5.7 highlights the ROC and Lift Curve for Neural Network. These are significant insights for our client due to the similarity and excellent efficiency. Next, the AUC for both Training and Validation are similar, with AUCs of 0.9704 and 0.9289, respectively. The lift curves meet at a point around 0.85 for the validation. Compared with the linear regression model and decision trees models above, the ROC values are higher, showing that the neural network boosted model would be a good focus for our client, as overfitting is curbed with the boosting and could provide an excellent model for focusing on the marine debris model prediction. To conclude, we also tried to run a model of two layers of neurons which can be found in the Appendix figure 6. We noticed that the neural network with two layers of 5 neurons produces a model with less accuracy. For example, the model with two layers has an R-square value of 0.4426 for validation which is a huge difference when compared to the model of one layer of 3 neurons and boosting. Moreover, the two layers model had 57 wrong predictions compared to 45 of the one-layer model with boosting.

Figure 5.6: Neural Network for Output Result

Model NTanH(3)NBoost(24)

Training

Y Binary

Measures	Value
Generalized RSquare	0.7480661
Entropy RSquare	0.6247285
RASE	0.2421129
Mean Abs Dev	0.1492609
Misclassification Rate	0.0745891
-LogLikelihood	167.84465
Sum Freq	791

Confusion Matrix

Actual Y Binary	Predicted Count Low	High
Low	570	21
High	38	162

Confusion Rates

Actual Y Binary	Predicted Rate Low	High
Low	0.964	0.036
High	0.190	0.810

Validation

Y Binary

Measures	Value
Generalized RSquare	0.6101745
Entropy RSquare	0.4791276
RASE	0.2877924
Mean Abs Dev	0.1862589
Misclassification Rate	0.1139241
-LogLikelihood	109.77604
Sum Freq	395

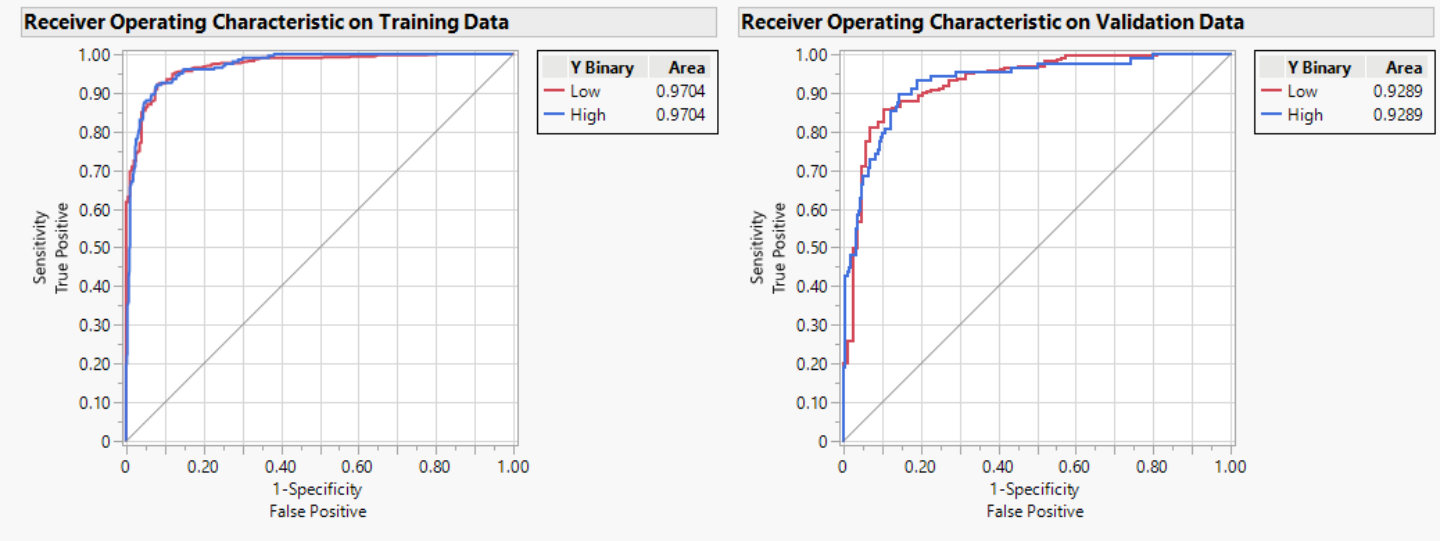
Confusion Matrix

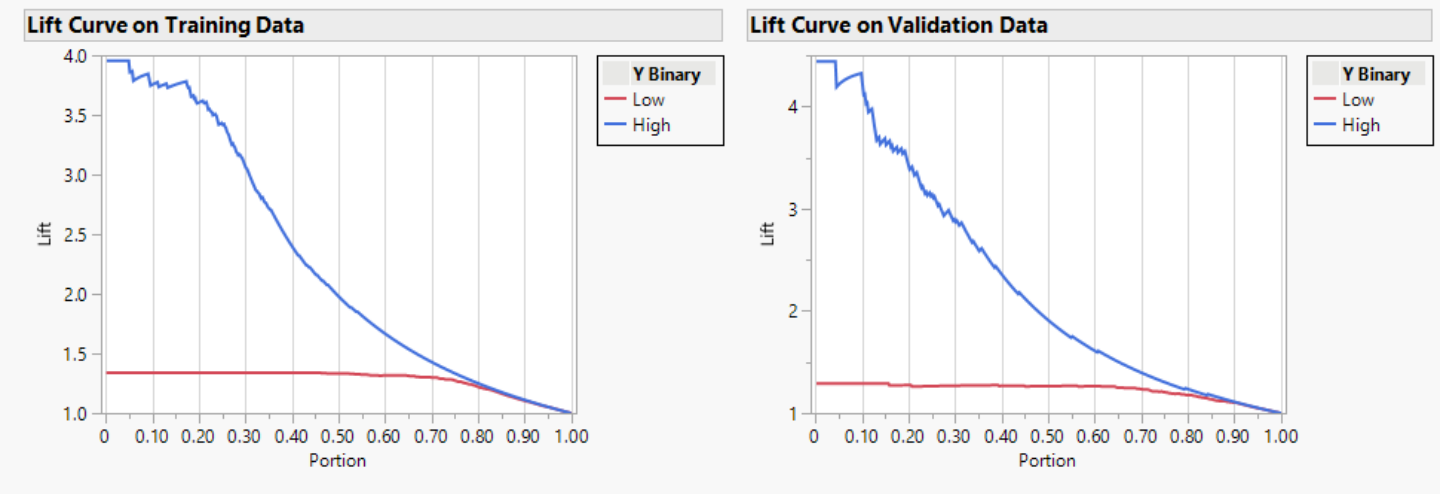
Actual Y Binary	Predicted Count Low	High
Low	291	15
High	30	59

Confusion Rates

Actual Y Binary	Predicted Rate Low	High
Low	0.951	0.049
High	0.337	0.663

Figure 5.7: ROC and Lift Curve for Neural Network





5.1.4 Support Vector Machine (SVM)

A support vector machine model using Y-Binary as the dependent variable was analyzed next. Through this analysis, the Best Model for the Comparison was Models 8. More specifically Model 8 has the Smallest Misclassification Rate of Threshold for Validation at 11.39%. Although the misclassification rate increased by about 0.03 from the training, this is still a very small change. This offers an opportunity for the client to trust that the SVM did a good job modeling and use its subsets to effectively focus on modeling instances of the marine debris data. Model 8 had the 0.4901 RSquare value of 0. Model 8 has a Cost was 1.62 and the Gamma was 0.22 and the number of SV was 426.

In addition, the confusion matrices in Figure 5.8 shows that some overfitting may have occurred. For example, the true positive in the training has a count of 586 and decreases to 296 in the validation. A decrease from training to validation can also be seen in the true negative. This could be due to the subset nature of SVM. Still, the counts are small compared to the size of the marine debris sample set and the training/validation predicted rates are somewhat similar. Hence, pointing to a good model for our client to focus on when using predictive modeling for marine debris variable comparison.

Figure 5.9 highlights the ROC and Lift Curve for SVM. The AUC for both Training and Validation are similar, with AUCs of 0.9912 and 0.9105, respectively. The lift curves meet at a point around 0.87 for the validation. Compared with the boosted neural networks model, the ROC values are very similar, showing the SVM would be worthwhile for our client to focus on for the marine debris model prediction.

Figure 5.8 Support Vector Machine for Model 8

Support Vector Machine Model 8

Model Summary

Response	Y Binary
Validation Method	Validation Column
Kernel Function	Radial Basis Function

Estimation Details

Cost	1.62
Gamma	0.21987

Measure	Training	Validation
Number of rows	791	395
Sum of Frequencies	791	395
Misclassification Rate	0.0518331	0.1139241
Number of Support Vectors	426	426

Confusion Matrix

Training

Actual Y Binary	Predicted Rate Low	Predicted Rate High
Low	0.990	0.010
High	0.175	0.825

Misclassification Rate: 0.0518

Actual Y Binary	Predicted Count Low	Predicted Count High
Low	585	6
High	35	165

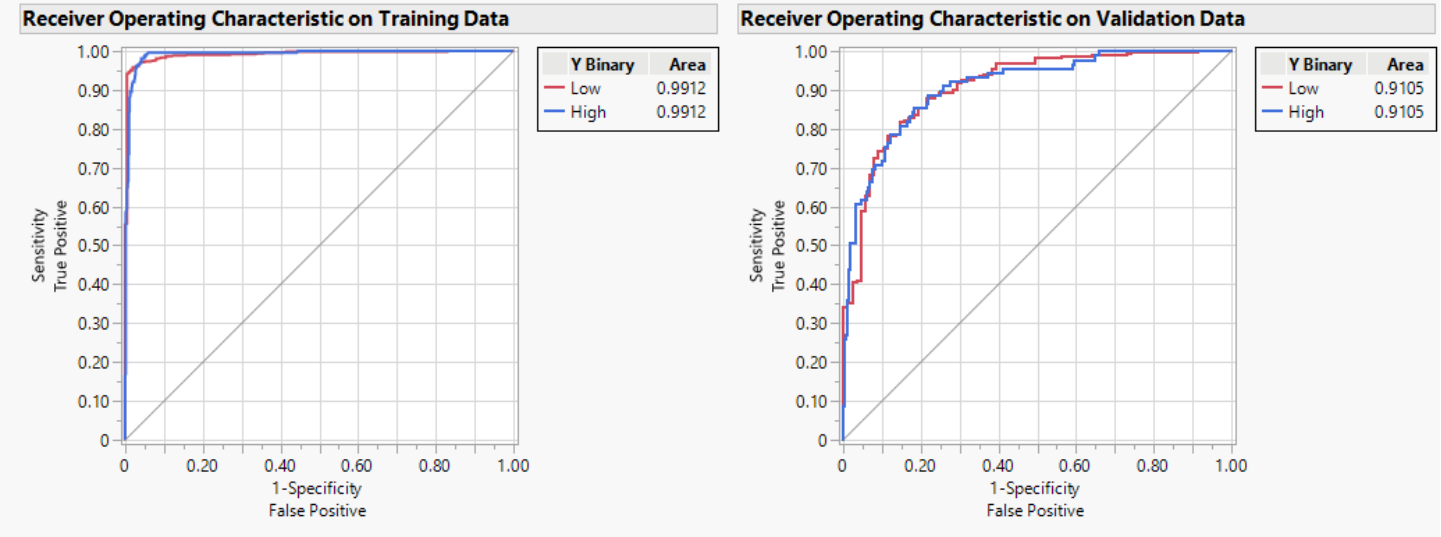
Validation

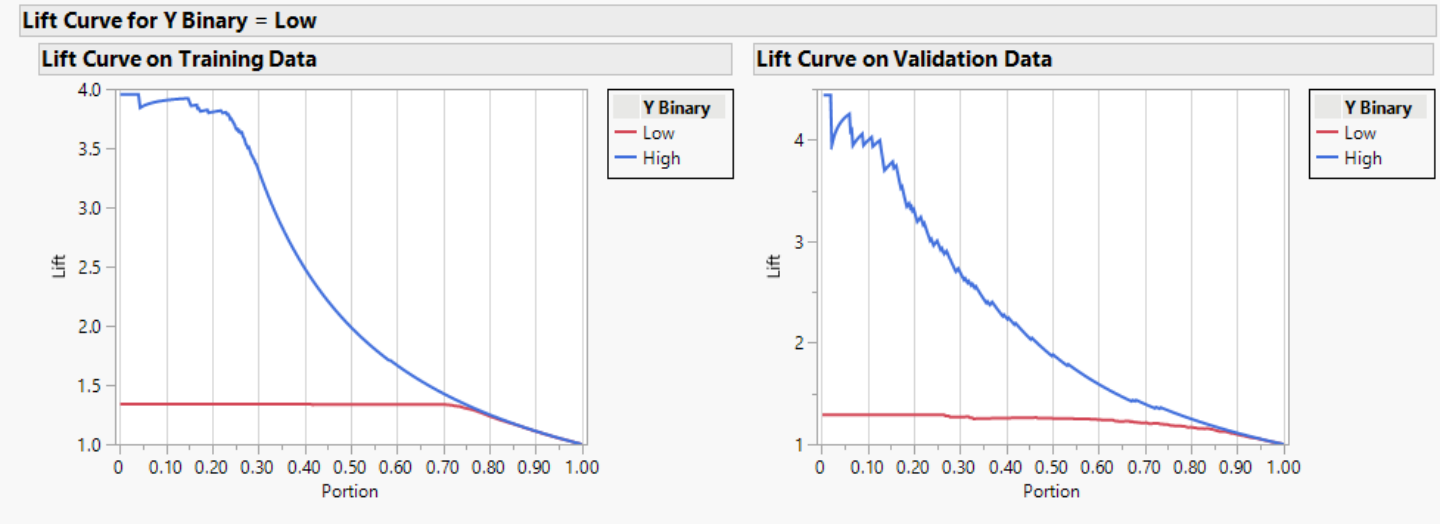
Actual Y Binary	Predicted Rate Low	Predicted Rate High
Low	0.967	0.033
High	0.393	0.607

Misclassification Rate: 0.1139

Actual Y Binary	Predicted Count Low	Predicted Count High
Low	296	10
High	35	54

Figure 5.9 ROC and Lift Curves for Support Vector Machine





5.1.5 Ensemble Modeling

After performing all analyses above, prediction formulas were saved for each model performed. Columns where the probability was “High” were labeled as DT, SVM, NN, and LR. Depicted in Figure 5.10, model’s DT and SVM contribute the most to the stacking ensemble since the LogWorth values are the highest and the PValues are the lowest. Focusing on the effect summary, the RSquare value is extremely high at 0.9757, meaning these models explain 97.57% of the data.

Decision tree modeling and SVM contribute most to the stacking ensemble. The LogWorth for the decision tree model is the most significant at 40.366 and 6.303 for the SVM. Both have significant PValues below 0.05. For this reason, we recommend that our client focuses on these two predictive modeling techniques to get the most out of pinpointing the types of marine debris and their significance.

Figure 5.10: Ensemble Modeling for Predictive Models

Effect Summary

Source	LogWorth	PValue
DT	40.366	0.00000
SVM	6.303	0.00000
NN	1.225	0.05953
LR	0.063	0.86409

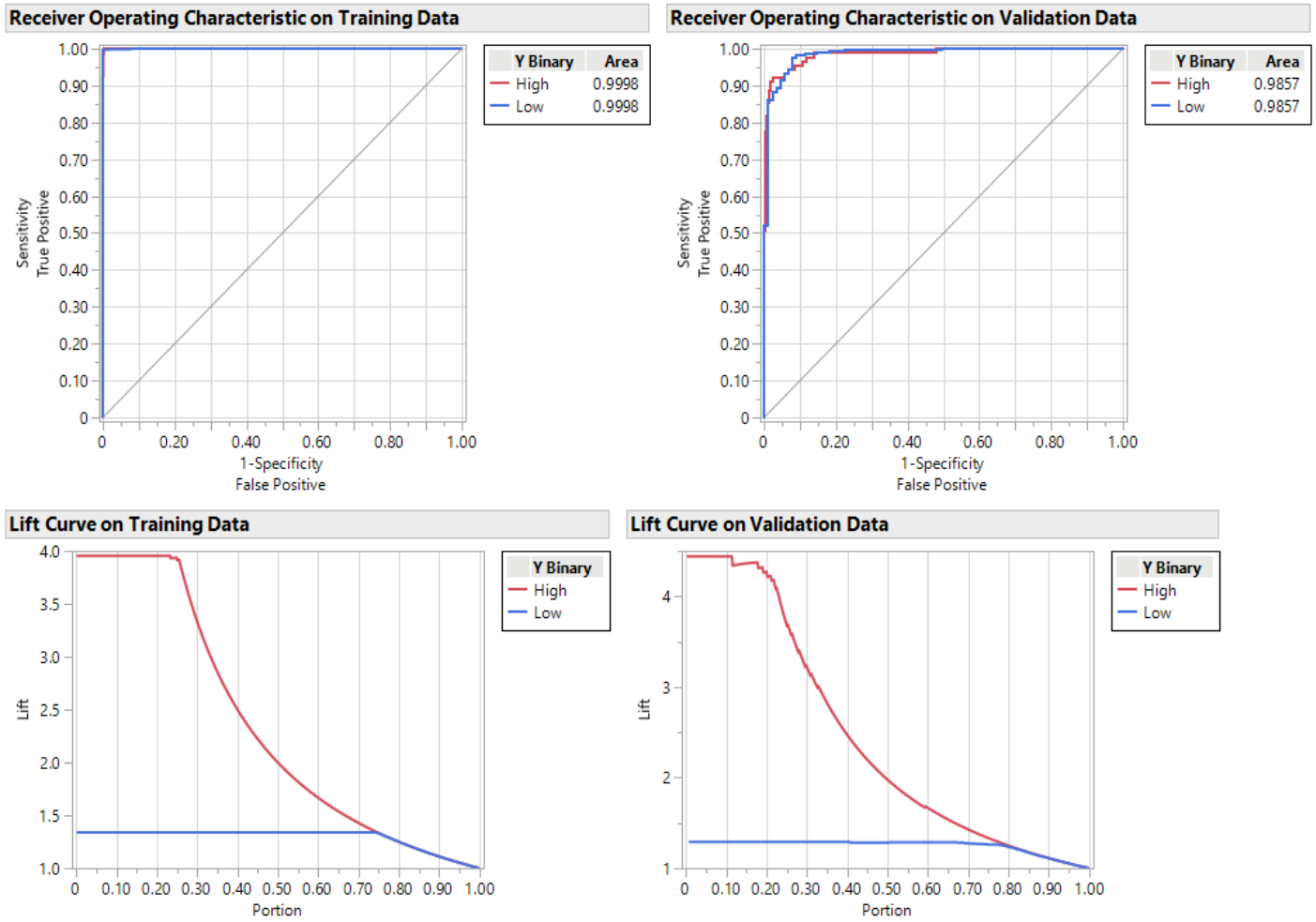
Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	436.39155	4	872.7831	<.0001*
Full	10.87040			
Reduced	447.26195			

RSquare (U)	0.9757
AICc	31.8172
BIC	55.1073
Observations (or Sum Wgts)	791

The ROC and Lift curves for the ensemble model show the highest AUC values of every model performed. The Training Model is almost at a perfect 1 value, and the line is almost completely in the top left corner, as shown in Figure 5.11. The Validation model has an AUC of 0.9857, which is also extremely high and close to the top left corner. These values mean that the model has predicted almost 100% true positives.

Figure 5.11: ROC and Lift curves for Ensemble model



5.2 Model Comparison Plots

We can see that the DT model has the best accuracy on both Training and Validation in Figure 5.12. On the other hand, LR has the least predictive accuracy which means that logistic regression should be our last option. Decision Trees can give us a clear understanding of which factors can result to High or Low amount of total debris step by step. As we address this business intelligence problem for our client, the Decision Tree analysis can also give us specific values that determine the final results. Such benchmarks could be used to predict patterns on locations that pinpoint high or low occurrences of debris pollution.

Figure 5.12: ROC Curves for Model Comparison Types

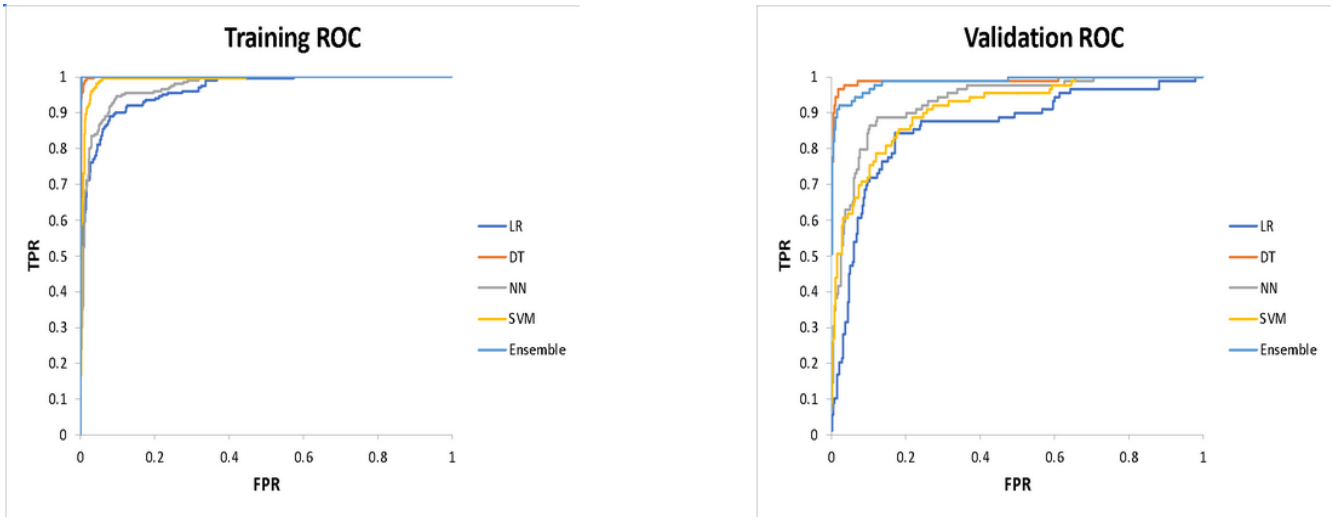
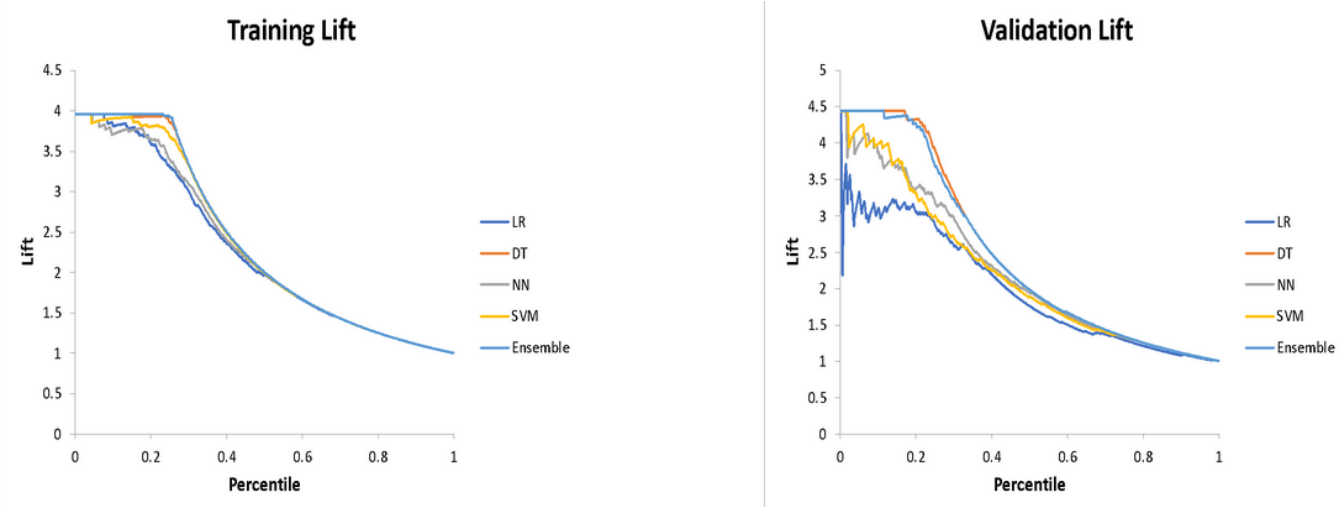


Figure 5.13 shows the different lift curves among each model type. Among all models, DT and SVM have the best lift curves. When analyzing the lift curves, DT has the least amount of overfitting. Between the Training and Validation DT curves, there is little variation between the two compared to the other models. This again shows the decision tree as the most accurate model for our dataset. The DT curve gives our client the best opportunity to predict the influencing factors within the dataset and guide decision making of areas to target cleanups.

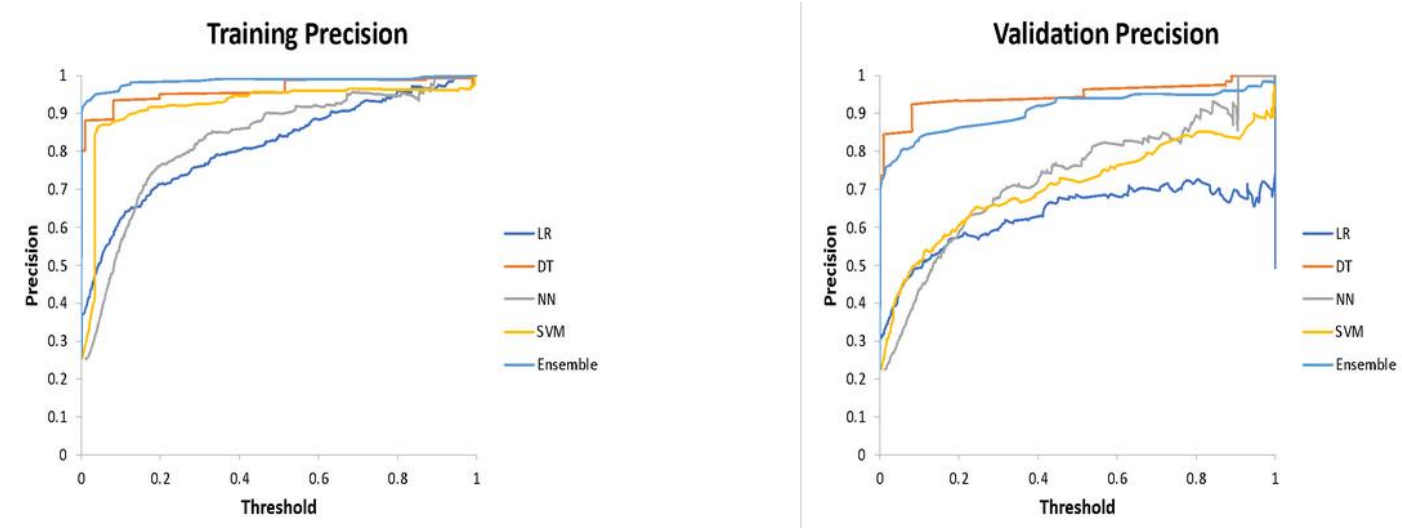
Figure 5.13: Lift Curves of Predictive Models



Precision curves can show us on Figure 5.14 that the precision of the model increases as the threshold increases too. Usually we tend to have the opposite scenario where the model has perfect precision in the beginning and as the precision decreases we catch more positive cases. Here we can see that again DT has

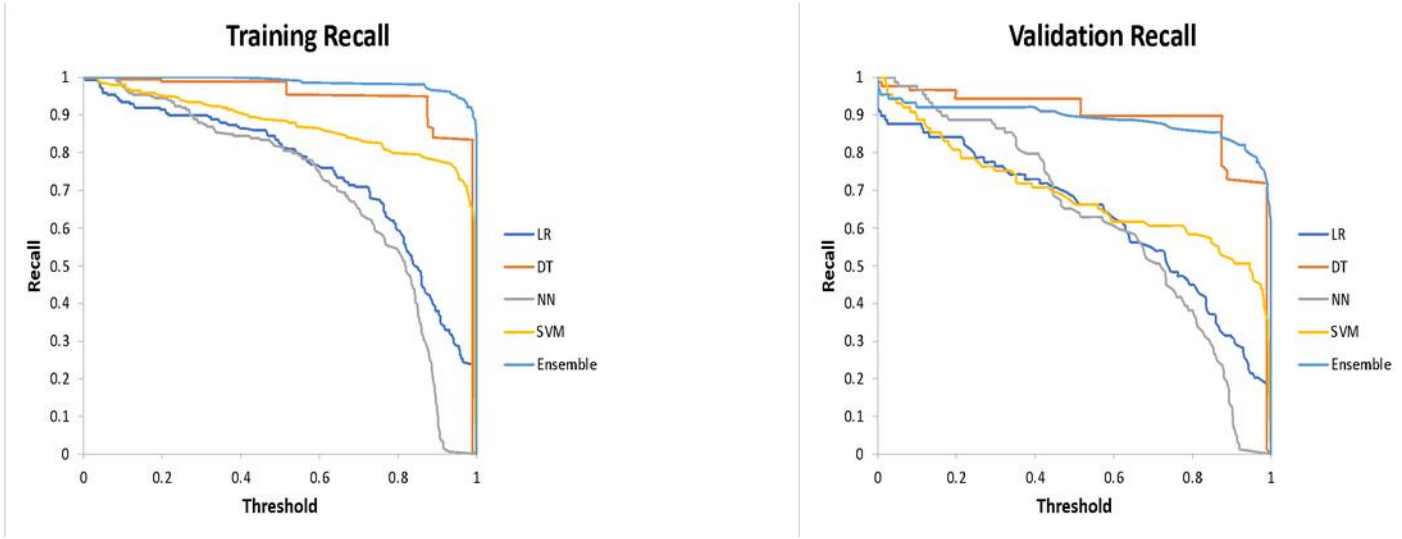
the highest precision and LR has the least precision for both Training and Validation. In this case, we would recommend the DT model. As the precision increases, the recall decreases. This would allow our client to effectively prioritize variables within the DT model and pinpoint these items when targeting marine debris. Hence, offering opportunities for our client to save money and direct clean up efforts appropriately.

Figure 5.14: Precision curves Model Comparison Types



Similar to the precision, the recall curves predict positive values and then show the probability that the value is actually true positive. DT has the highest accuracy and NN has the lowest prediction. As a result, we can also guide our client to review the recall in Figure 5.15 and gauge opportunities to target locations with high amounts of marine debris (positive case). This will allow our client to again maximize their resources and increase the speed of the cleanups/target a higher volume of areas for cleanups.

Figure 5.15: Recall Curves of Predictive Models

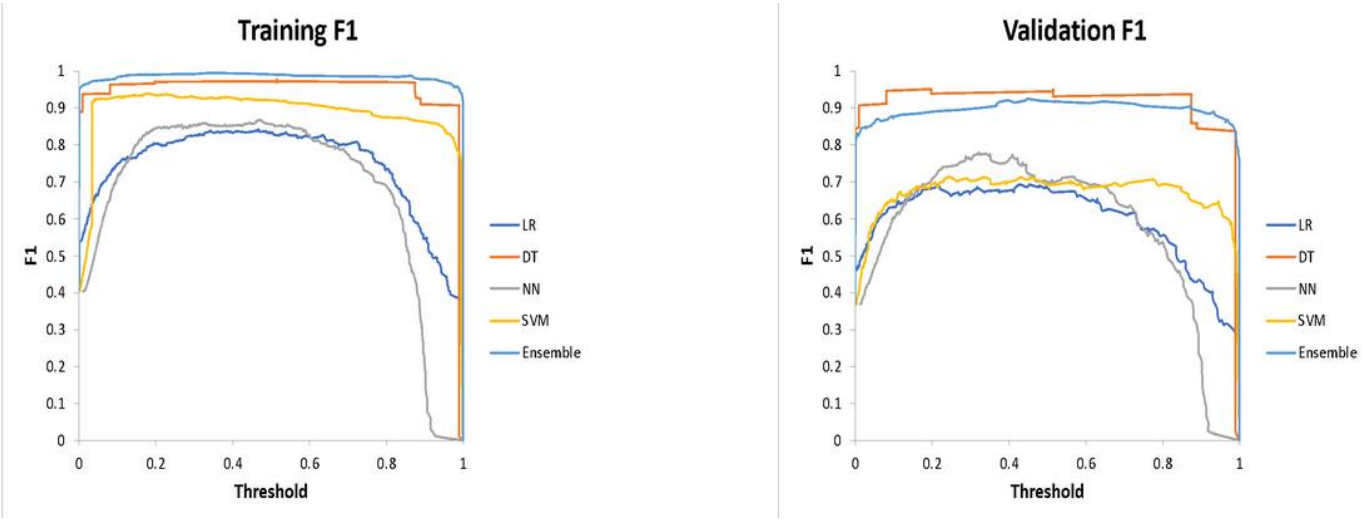


In regards to the F1 curves, DT has the most harmonic F-score in comparison to the other models. It can be noted in Figure 5.16, that DT has the most area underneath its curve. Larger the areas under the F1 curve implies higher prediction capability of the specific model.

This is beneficial for our client because the DT curve again will provide an opportunity to predict the best marine debris variables to focus on. The F1 score can help our client better understand which analysis is the best to focus on and give us the most accurate predictions for pinpointing marine debris variables within the dataset.

In addition, the boosted neural network and LR curves are the worst model to focus on because of the lack of area underneath their curves. Looking at the summary Figure the percentages of F1 underneath LR and NN for both training and validation have the smallest percentages. Hence pointing to less accuracy of these models. For this reason, we would not recommend these curves for our client. Both of these curves would most likely yield lower opportunities to capture marine debris and not maximize the client's resources.

Figure 5.16: F1 Curves of Predictive Models



5.3 Summary Table

Looking at the Summary Table in Table 5.17, we can analyze the results of each performance measure used easily. Performance measures for each model include F1, Precision, Recall, and AUC. Worst values are highlighted in red and bolded, while highest values are highlighted in green and bolded. The Table clearly shows which model(s) should be used for predictive modeling.

Table 5.17: Summary Table of Predictive Models

Training	LR	DT	NN	SVM	Ensemble
F1-thd	0.4670	0.5159	0.4732	0.1798	0.3449
F1	84.03%	97.78%	86.75%	93.89%	99.50%
Precision	82.61%	96.59%	90.27%	91.87%	99.01%
Recall	85.50%	99.00%	83.50%	96.00%	100.00%
AUC	96.20%	99.81%	97.22%	99.12%	99.98%

Validation	LR	DT	NN	SVM	Ensemble
F1-thd	0.4518	0.1984	0.3226	0.2423	0.4484
F1	69.23%	95.03%	77.78%	71.43%	92.57%
Precision	67.74%	93.48%	70.64%	65.42%	94.19%
Recall	70.79%	96.63%	86.52%	78.65%	91.01%
AUC	85.78%	99.00%	92.96%	91.05%	98.57%

The precision, recall, and F1 results are based on the threshold with the best F1 results in respective Training and Validation Table. In the Training Table, it is clear that logistic regression is the worst model to use in predicting our dataset, as this measure had more performance measure results that were lower than the others. Alternatively, the ensemble model resulted in the best model to use, as all highest values were output for this model.

In the Validation Table, logistic regression is again the model with the lowest values. Since logistic regression offered the same results in both Training and Validation columns, logistic regression should not be used in predictive modeling for our dataset. Decision trees contained the most amount of high values. Comparing the two, it is clear that decision trees offer some of the highest performance measure results out of each model analyzed. In the Training Table, ensemble modeling had the highest results, but decision trees were a close second. Using this logic, we believe it is clear that decision tree modeling should be used in predicting the probability of positive outcomes in our dataset. In relation to providing insights for our client, the best results for performance can be found using the ensemble modeling and decision tree methods. Again, the decision tree allows for the best precision, it splits the counts of states and then pinpoints variables such as season and types of debris found in these areas (i.e. debris behind barrier and survey year). This is beneficial for our client because it will allow them to increase their resources within specific seasons and locations to prevent overspending. Ultimately, our client can use the DT and ensemble models to analyze instances and map out how their resources could be used in years to come. Hence, providing a strategic advantage for the company's marine debris cleanup efforts.

6. Conclusions, Discussion and Recommendations

We are confident that our findings regarding marine debris pollution will be an asset for many businesses that wish to clean up marine debris within polluted waters. Due to its widespread impact on the world, marine debris is impacting a multitude of marine ecosystems and human communities. From this analysis we have a number of conclusions, discussion items and recommendations.

6.1 Conclusion

In conclusion, this marine debris analysis provided a number of valuable insights for helping solve our business intelligence problem. The hard plastics were most prevalent among all other wastes.This can be seen in modeling such as decision tree and ensemble predictive modeling.

Second, organizations should focus their efforts on the Gulf of Mexico and Hawaii.These were areas that commonly appeared on the decision tree modeling. Decision trees should be used more frequently over other forms of analyses, as this model had the most accuracy compared to the other predictive models. We are confident that this study will effectively allow clients to capture marine debris from the polluted water more efficiently.

6.2 Discussion

This analysis of marine debris attempted to find the correlation between the season and the total number of the debris, however, not any major connection between those variables was found.

Throughout this analysis, Decision Tree analysis was an effective tool for figuring out influential factors of marine debris. The locations provided with high marine debris totals, such as Texas and Hawaii, provided valuable insights. These could be areas where ocean currents carry marine debris to specific locations.

6.3 Recommendations

Based on our study, we have a number of recommendations for the further analysis of marine debris. This type of data can help to provide an in-depth understanding of marine debris sources and provide real-world, data-driven insights.

Based on the findings of the Multiple regression analysis, we recommend the Government authorities to regulate the disposal of common marine debris such as Hard Plastics that can increase the micro plastic pollution in the water. The collection of micro pollution can be hard to collect and can also be harmful for human health; therefore, the best way to face such a problem would be to create more strict regulation that would prevent the spread of such pollution.

Based on the ANOVA analysis we were able to explore which states have the highest number of plastic marine pollution. For example, in Hawaii and Texas it is more common to have higher amounts of marine pollution, which could imply the necessity of cleanup operations that would make sure that the ocean is a healthy and clean habitat for the marine ecosystem.

Furthermore, the Scatter Plot analysis was able to give us more information about which types of plastic pollution is more frequently found in the waters of these locations. Such information can be important before a cleanup operation since it is known which techniques would be more efficient and would result in a faster cleansing of the polluted waters.

Appendix A
Figure 1: Scatterplot Matrix

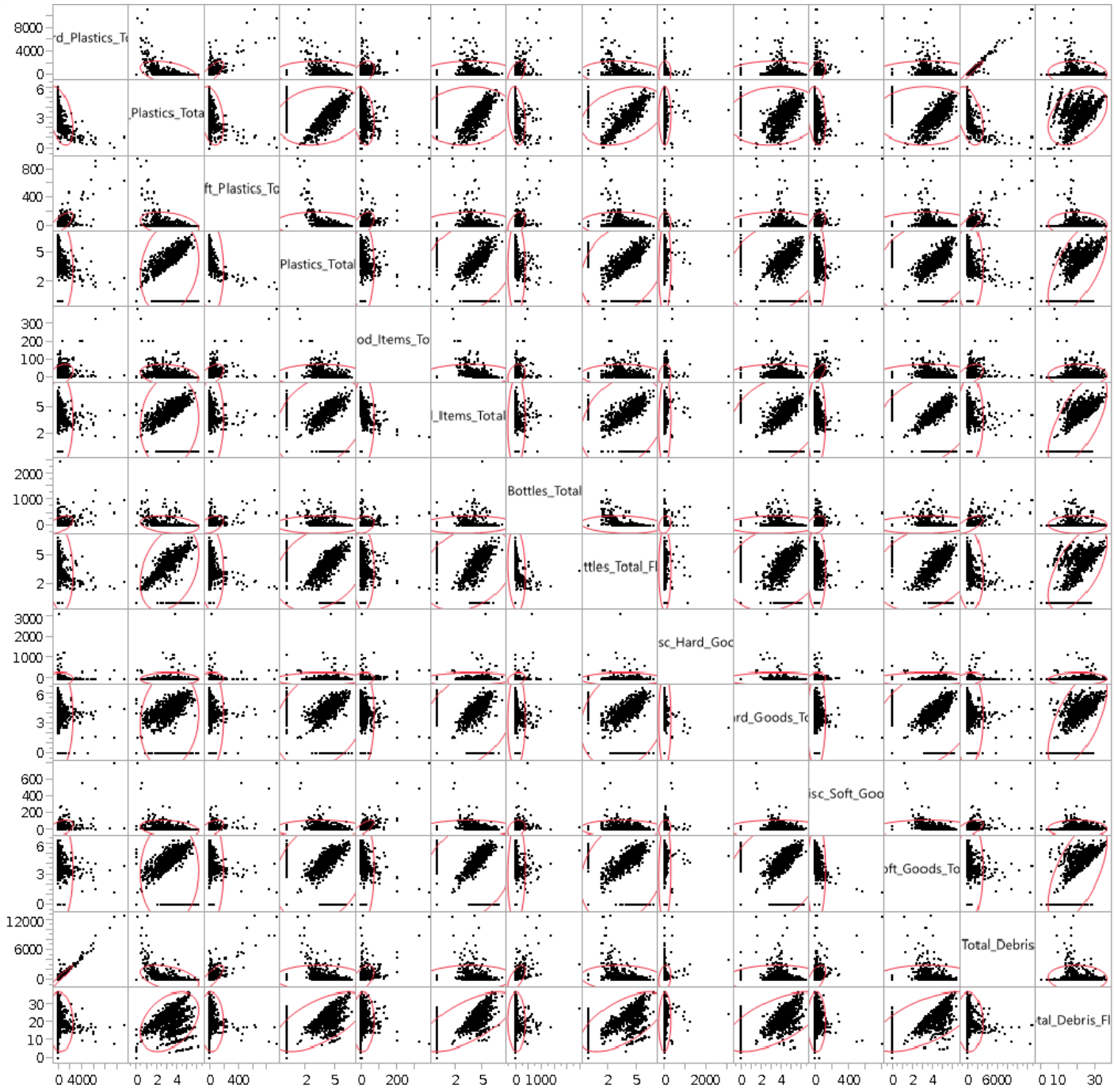


Figure 2: Scatterplot Matrix Pairwise Correlations

Pairwise Correlations

Variable	by Variable	Correlation	Count	Lower 95%	Upper 95%	Signif Prob
Total_Debris	Hard_Plastics_Total	0.9775	1186	0.9749	0.9799	<.0001*
Total_Debris	Soft_Plastics_Total	0.8052	1186	0.7842	0.8243	<.0001*
Misc_Soft_Goods	Food_Items_Total	0.7834	1186	0.7604	0.8044	<.0001*
Soft_Plastics_Total	Hard_Plastics_Total	0.7697	1186	0.7454	0.7919	<.0001*
Total_Debris_Flux	Food_Items_Total_Flux	0.6721	1186	0.6396	0.7022	<.0001*
Total_Debris	Bottles_Total	0.6467	1186	0.6124	0.6787	<.0001*
Total_Debris_Flux	Misc_Hard_Goods_Total_Flux	0.6358	1186	0.6006	0.6685	<.0001*
Total_Debris_Flux	Soft_Plastics_Total_Flux	0.6312	1186	0.5957	0.6642	<.0001*
Total_Debris_Flux	Misc_Soft_Goods_Total_Flux	0.6285	1186	0.5928	0.6618	<.0001*
Total_Debris_Flux	Bottles_Total_Flux	0.6259	1186	0.5899	0.6593	<.0001*
Misc_Soft_Goods	Soft_Plastics_Total	0.5668	1186	0.5268	0.6042	<.0001*
Total_Debris	Misc_Soft_Goods	0.5504	1186	0.5094	0.5889	<.0001*
Bottles_Total	Hard_Plastics_Total	0.5294	1186	0.4871	0.5691	<.0001*
Food_Items_Total	Soft_Plastics_Total	0.5257	1186	0.4833	0.5657	<.0001*
Total_Debris	Food_Items_Total	0.5059	1186	0.4623	0.5471	<.0001*
Misc_Soft_Goods	Hard_Plastics_Total	0.4831	1186	0.4383	0.5256	<.0001*
Hard_Plastics_Total_Flux	Hard_Plastics_Total	-0.4608	1186	-0.5045	-0.4147	<.0001*
Bottles_Total	Soft_Plastics_Total	0.4594	1186	0.4133	0.5031	<.0001*
Total_Debris	Hard_Plastics_Total_Flux	-0.4546	1186	-0.4986	-0.4082	<.0001*
Food_Items_Total	Hard_Plastics_Total	0.4393	1186	0.3921	0.4841	<.0001*
Total_Debris_Flux	Hard_Plastics_Total_Flux	0.3968	1186	0.3477	0.4437	<.0001*
Bottles_Total_Flux	Hard_Plastics_Total_Flux	0.3408	1186	0.2895	0.3901	<.0001*
Soft_Plastics_Total	Hard_Plastics_Total_Flux	-0.3391	1186	-0.3885	-0.2877	<.0001*
Misc_Hard_Goods_Total_Flux	Food_Items_Total_Flux	0.3372	1186	0.2857	0.3867	<.0001*
Misc_Soft_Goods_Total_Flux	Food_Items_Total_Flux	0.3117	1186	0.2594	0.3622	<.0001*
Misc_Soft_Goods_Total_Flux	Misc_Hard_Goods_Total_Flux	0.3109	1186	0.2586	0.3614	<.0001*
Bottles_Total_Flux	Soft_Plastics_Total_Flux	0.2916	1186	0.2386	0.3428	<.0001*
Food_Items_Total	Hard_Plastics_Total_Flux	-0.2842	1186	-0.3357	-0.2311	<.0001*
Bottles_Total	Hard_Plastics_Total_Flux	-0.2793	1186	-0.3310	-0.2260	<.0001*
Misc_Soft_Goods	Bottles_Total	0.2775	1186	0.2241	0.3292	<.0001*
Misc_Soft_Goods	Hard_Plastics_Total_Flux	-0.2770	1186	-0.3287	-0.2236	<.0001*
Food_Items_Total_Flux	Soft_Plastics_Total_Flux	0.2738	1186	0.2203	0.3256	<.0001*
Bottles_Total_Flux	Food_Items_Total_Flux	0.2738	1186	0.2203	0.3256	<.0001*
Bottles_Total	Food_Items_Total	0.2721	1186	0.2186	0.3240	<.0001*
Misc_Hard_Goods	Bottles_Total	0.2681	1186	0.2145	0.3202	<.0001*
Misc_Hard_Goods_Total_Flux	Soft_Plastics_Total_Flux	0.2574	1186	0.2034	0.3097	<.0001*
Misc_Soft_Goods_Total_Flux	Bottles_Total_Flux	0.2446	1186	0.1903	0.2974	<.0001*
Misc_Soft_Goods_Total_Flux	Soft_Plastics_Total_Flux	0.2408	1186	0.1864	0.2937	<.0001*
Total_Debris	Misc_Hard_Goods	0.2230	1186	0.1682	0.2764	<.0001*
Misc_Hard_Goods_Total_Flux	Bottles_Total_Flux	0.1966	1186	0.1413	0.2507	<.0001*
Soft_Plastics_Total_Flux	Hard_Plastics_Total_Flux	0.1930	1186	0.1376	0.2472	<.0001*
Misc_Soft_Goods	Misc_Hard_Goods	0.1814	1186	0.1258	0.2359	<.0001*
Misc_Hard_Goods	Food_Items_Total	0.1654	1186	0.1095	0.2203	<.0001*
Bottles_Total_Flux	Bottles_Total	-0.1521	1186	-0.2072	-0.0960	<.0001*
Total_Debris	Bottles_Total_Flux	-0.1308	1186	-0.1863	-0.0744	<.0001*
Misc_Hard_Goods_Total_Flux	Hard_Plastics_Total_Flux	0.1244	1186	0.0680	0.1801	<.0001*
Bottles_Total_Flux	Hard_Plastics_Total	-0.1242	1186	-0.1799	-0.0678	<.0001*
Food_Items_Total_Flux	Hard_Plastics_Total_Flux	0.1046	1186	0.0480	0.1606	0.0003*
Misc_Hard_Goods	Soft_Plastics_Total	0.0999	1186	0.0433	0.1560	0.0006*
Misc_Hard_Goods	Food_Items_Total_Flux	0.0999	1186	0.0432	0.1559	0.0006*
Bottles_Total_Flux	Soft_Plastics_Total	-0.0912	1186	-0.1474	-0.0345	0.0017*
Misc_Hard_Goods_Total_Flux	Food_Items_Total	0.0851	1186	0.0283	0.1413	0.0034*
Bottles_Total	Food_Items_Total_Flux	0.0805	1186	0.0237	0.1368	0.0056*
Total_Debris	Food_Items_Total_Flux	0.0794	1186	0.0226	0.1357	0.0062*
Misc_Soft_Goods_Total_Flux	Food_Items_Total	0.0783	1186	0.0214	0.1346	0.0070*
Misc_Soft_Goods_Total_Flux	Misc_Hard_Goods	0.0775	1186	0.0206	0.1338	0.0076*
Misc_Soft_Goods_Total_Flux	Bottles_Total	0.0766	1186	0.0197	0.1329	0.0083*
Total_Debris_Flux	Misc_Hard_Goods	0.0742	1186	0.0174	0.1306	0.0105*
Total_Debris	Misc_Soft_Goods_Total_Flux	0.0742	1186	0.0173	0.1306	0.0106*
Food_Items_Total_Flux	Food_Items_Total	0.0726	1186	0.0158	0.1290	0.0123*
Total_Debris_Flux	Hard_Plastics_Total	-0.0726	1186	-0.1290	-0.0158	0.0124*
Misc_Hard_Goods	Hard_Plastics_Total	0.0652	1186	0.0083	0.1216	0.0248*
Food_Items_Total_Flux	Soft_Plastics_Total	0.0633	1186	0.0064	0.1197	0.0294*
Total_Debris	Misc_Hard_Goods_Total_Flux	0.0624	1186	0.0055	0.1189	0.0318*
Soft_Plastics_Total_Flux	Soft_Plastics_Total	-0.0609	1186	-0.1174	-0.0039	0.0361*
Food_Items_Total_Flux	Hard_Plastics_Total	0.0598	1186	0.0029	0.1164	0.0394*
Misc_Soft_Goods_Total_Flux	Hard_Plastics_Total	0.0585	1186	0.0016	0.1150	0.0440*
Total_Debris_Flux	Total_Debris	-0.0575	1186	-0.1141	-0.0006	0.0476*
Misc_Soft_Goods	Misc_Hard_Goods_Total_Flux	0.0563	1186	-0.0007	0.1128	0.0528
Misc_Soft_Goods	Bottles_Total_Flux	-0.0562	1186	-0.1127	0.0008	0.0532
Misc_Hard_Goods_Total_Flux	Hard_Plastics_Total	0.0559	1186	-0.0011	0.1124	0.0545
Misc_Hard_Goods_Total_Flux	Soft_Plastics_Total	0.0554	1186	-0.0015	0.1120	0.0564
Misc_Hard_Goods_Total_Flux	Bottles_Total	0.0543	1186	-0.0027	0.1109	0.0617
Misc_Soft_Goods	Food_Items_Total_Flux	0.0529	1186	-0.0041	0.1095	0.0688
Misc_Hard_Goods	Soft_Plastics_Total_Flux	0.0525	1186	-0.0044	0.1091	0.0706
Total_Debris_Flux	Soft_Plastics_Total	-0.0507	1186	-0.1073	0.0062	0.0809
Misc_Soft_Goods_Total_Flux	Misc_Soft_Goods	0.0505	1186	-0.0064	0.1071	0.0820
Misc_Soft_Goods_Total_Flux	Soft_Plastics_Total	0.0474	1186	-0.0095	0.1040	0.1028
Soft_Plastics_Total_Flux	Hard_Plastics_Total	-0.0414	1186	-0.0981	0.0155	0.1540
Bottles_Total_Flux	Food_Items_Total	-0.0351	1186	-0.0919	0.0218	0.2268
Total_Debris	Soft_Plastics_Total_Flux	-0.0308	1186	-0.0876	0.0261	0.2887
Misc_Soft_Goods_Total_Flux	Hard_Plastics_Total_Flux	0.0275	1186	-0.0294	0.0843	0.3433
Food_Items_Total	Soft_Plastics_Total_Flux	0.0260	1186	-0.0310	0.0828	0.3707
Total_Debris_Flux	Bottles_Total	-0.0248	1186	-0.0816	0.0321	0.3930
Total_Debris_Flux	Food_Items_Total	0.0196	1186	-0.0374	0.0764	0.4999
Total_Debris_Flux	Misc_Soft_Goods	-0.0136	1186	-0.0705	0.0434	0.6402
Misc_Hard_Goods_Total_Flux	Misc_Hard_Goods	0.0117	1186	-0.0453	0.0686	0.6877
Misc_Soft_Goods	Soft_Plastics_Total_Flux	0.0052	1186	-0.0517	0.0621	0.8568
Misc_Hard_Goods	Bottles_Total_Flux	0.0036	1186	-0.0534	0.0605	0.9021
Misc_Hard_Goods	Hard_Plastics_Total_Flux	0.0021	1186	-0.0548	0.0591	0.9413
Bottles_Total	Soft_Plastics_Total_Flux	0.0015	1186	-0.0555	0.0584	0.9599

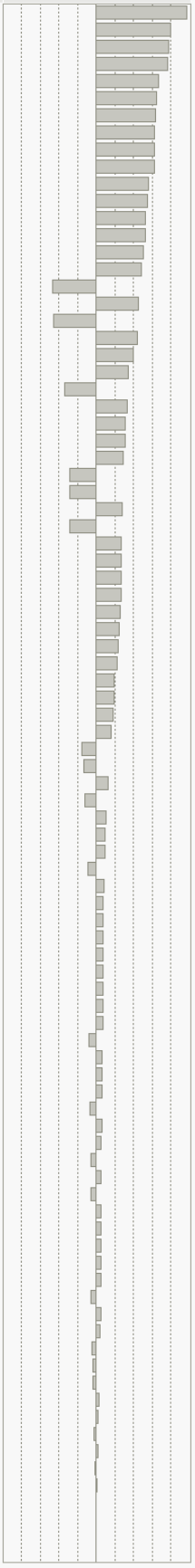


Figure 3: Scatterplot RSquare Values

Correlations

	Hard_Plastics_Total	Hard_Plastics_Total_Flux	Soft_Plastics_Total	Soft_Plastics_Total_Flux	Food_Items_Total	Food_Items_Total_Flux	Bottles_Total	Bottles_Total_Flux	Misc_Hard_Goods	Misc_Hard_Goods_Total_Flux	Misc_Soft_Goods	Misc_Soft_Goods_Total_Flux	Total_Debris	Total_Debris_Flux
Hard_Plastics_Total	1.0000	-0.4608	0.7697	-0.0414	0.4393	0.0598	0.5294	-0.1242	0.0652	0.0559	0.4831	0.0585	0.9775	-0.0726
Hard_Plastics_Total_Flux	-0.4608	1.0000	-0.3391	0.1930	-0.2842	0.1046	-0.2793	0.3408	0.0021	0.1244	-0.2770	0.0275	-0.4546	0.3968
Soft_Plastics_Total	0.7697	-0.3391	1.0000	-0.0609	0.5257	0.0633	0.4594	-0.0912	0.0999	0.0554	0.5668	0.0474	0.8052	-0.0507
Soft_Plastics_Total_Flux	-0.0414	0.1930	-0.0609	1.0000	0.0260	0.2738	0.0015	0.2916	0.0525	0.2574	0.0052	0.2408	-0.0308	0.6312
Food_Items_Total	0.4393	-0.2842	0.5257	0.0260	1.0000	0.0726	0.2721	-0.0351	0.1654	0.0851	0.7834	0.0783	0.5059	0.0196
Food_Items_Total_Flux	0.0598	0.1046	0.0633	0.2738	0.0726	1.0000	0.0805	0.2738	0.0999	0.3372	0.0529	0.3117	0.0794	0.6721
Bottles_Total	0.5294	-0.2793	0.4594	0.0015	0.2721	0.0805	1.0000	-0.1521	0.2681	0.0543	0.2775	0.0766	0.6467	-0.0248
Bottles_Total_Flux	-0.1242	0.3408	-0.0912	0.2916	-0.0351	0.2738	-0.1521	1.0000	0.0036	0.1966	-0.0562	0.2446	-0.1308	0.6259
Misc_Hard_Goods	0.0652	0.0021	0.0999	0.0525	0.1654	0.0999	0.2681	0.0036	1.0000	0.0117	0.1814	0.0775	0.2230	0.0742
Misc_Hard_Goods_Total_Flux	0.0559	0.1244	0.0554	0.2574	0.0851	0.3372	0.0543	0.1966	0.0117	1.0000	0.0563	0.3109	0.0624	0.6358
Misc_Soft_Goods	0.4831	-0.2770	0.5668	0.0052	0.7834	0.0529	0.2775	-0.0562	0.1814	0.0563	1.0000	0.0505	0.5504	-0.0136
Misc_Soft_Goods_Total_Flux	0.0585	0.0275	0.0474	0.2408	0.0783	0.3117	0.0766	0.2446	0.0775	0.3109	0.0505	1.0000	0.0742	0.6285
Total_Debris	0.9775	-0.4546	0.8052	-0.0308	0.5059	0.0794	0.6467	-0.1308	0.2230	0.0624	0.5504	0.0742	1.0000	-0.0575
Total_Debris_Flux	-0.0726	0.3968	-0.0507	0.6312	0.0196	0.6721	-0.0248	0.6259	0.0742	0.6358	-0.0136	0.6285	-0.0575	1.0000

The correlations are estimated by Row-wise method.

Figure 4: Model depicting Back Barrier (Slatt, 2013)

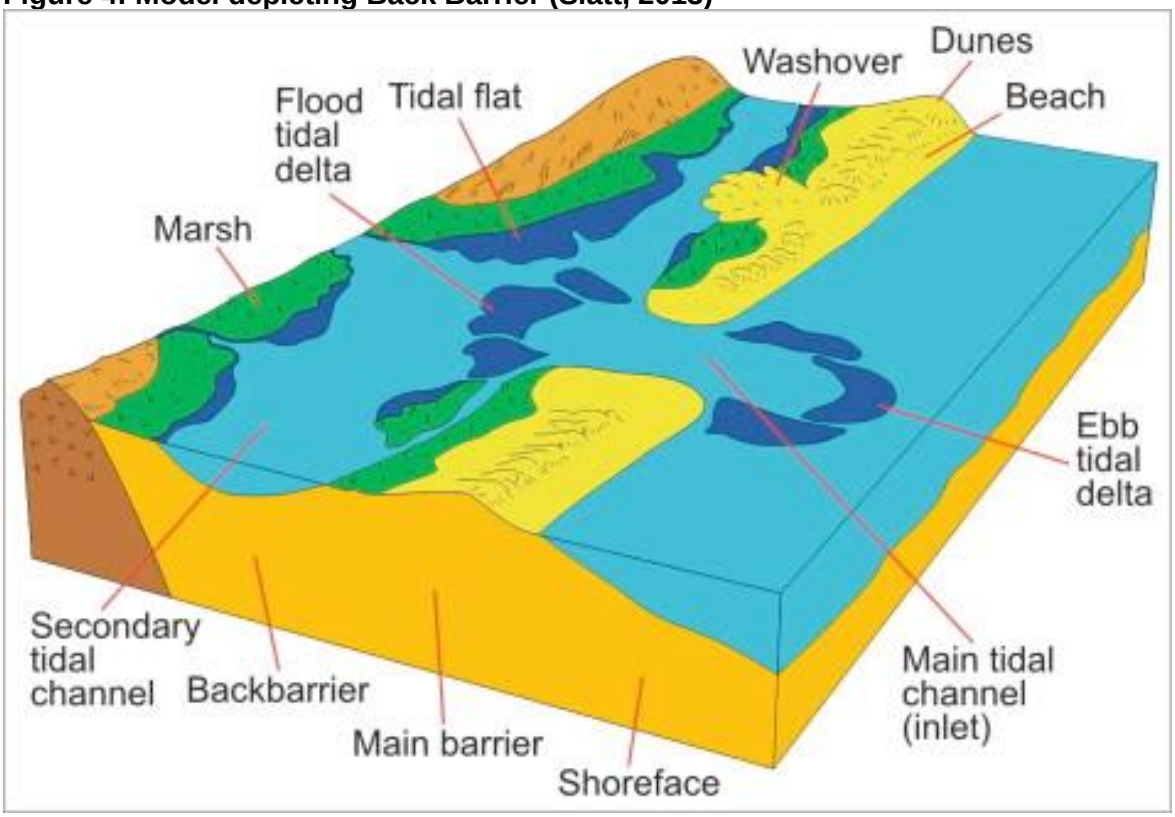


Figure 5: Decision Tree Full View

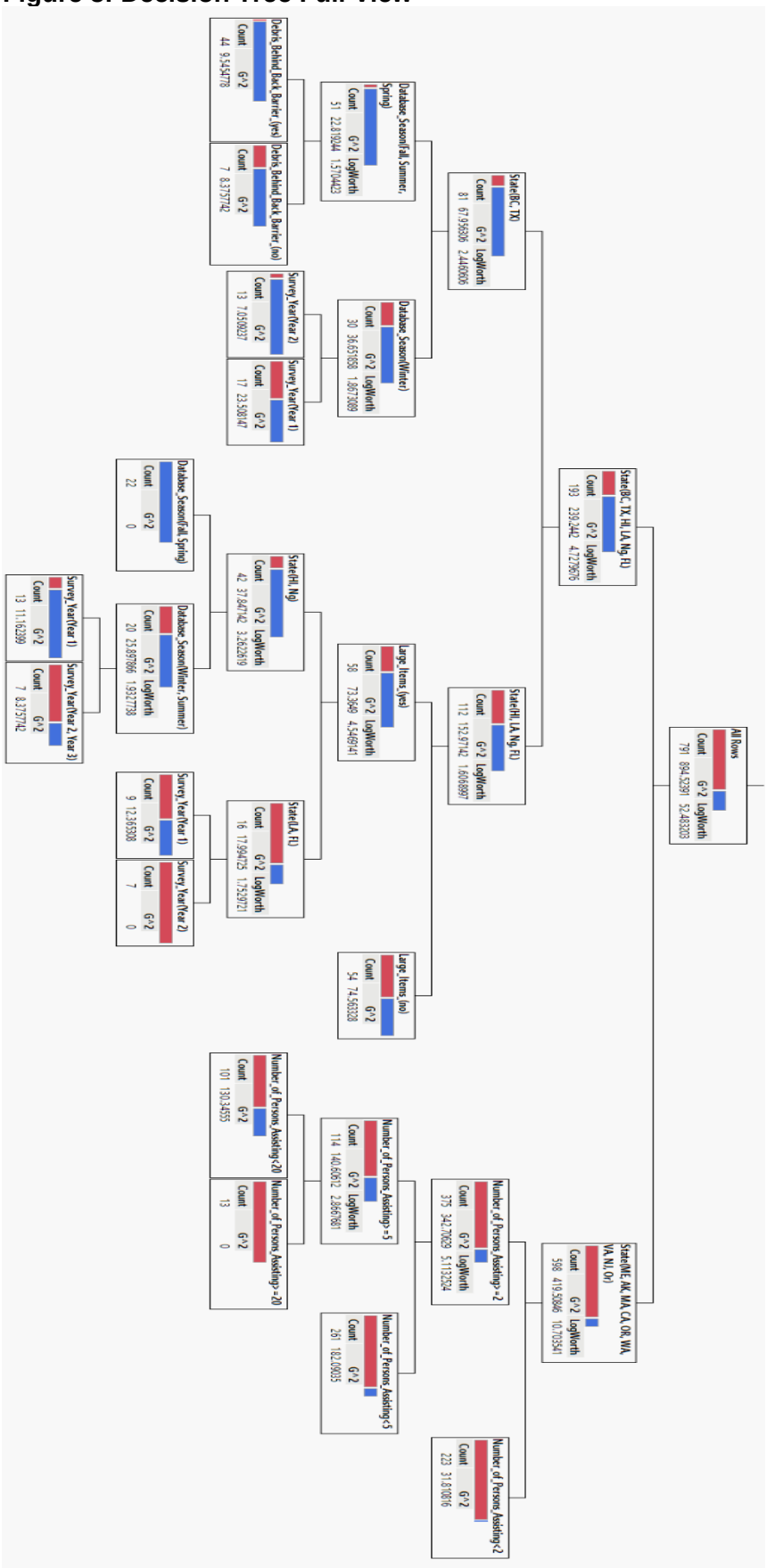
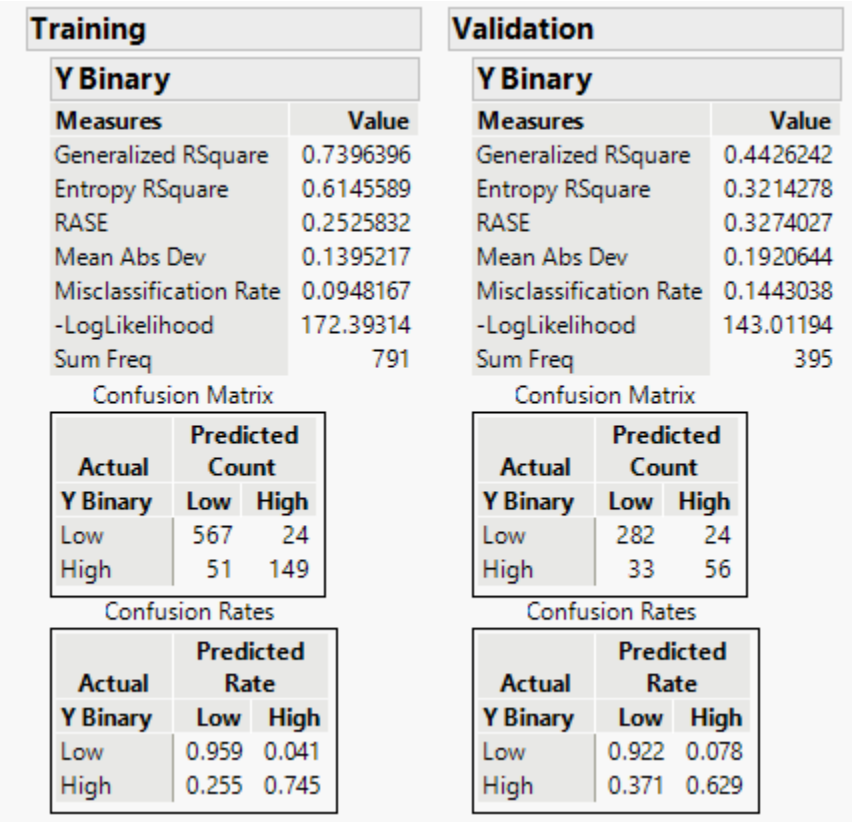


Figure 6: Neural Network with 2 layers



Training

Y Binary

Measures	Value
Generalized RSquare	0.7396396
Entropy RSquare	0.6145589
RASE	0.2525832
Mean Abs Dev	0.1395217
Misclassification Rate	0.0948167
-LogLikelihood	172.39314
Sum Freq	791

Confusion Matrix

Actual	Predicted Count	
Y Binary	Low	High
Low	567	24
High	51	149

Confusion Rates

Actual	Predicted Rate	
Y Binary	Low	High
Low	0.959	0.041
High	0.255	0.745

Validation

Y Binary

Measures	Value
Generalized RSquare	0.4426242
Entropy RSquare	0.3214278
RASE	0.3274027
Mean Abs Dev	0.1920644
Misclassification Rate	0.1443038
-LogLikelihood	143.01194
Sum Freq	395

Confusion Matrix

Actual	Predicted Count	
Y Binary	Low	High
Low	282	24
High	33	56

Confusion Rates

Actual	Predicted Rate	
Y Binary	Low	High
Low	0.922	0.078
High	0.371	0.629

Figure 7: Support Vector Machine Model Comparison

Show	Method	Kernel Function	Cost	Gamma	# SV	Training Misclassification Rate	Validation Misclassification Rate	Validation Generalized RSquare	Probability Threshold	Threshold Validation Misclassification Rate	Best
☐	Model 1	Radial Basis Function	2.00413	0.46125	583	0.01643	0.13671	0.49616	0.50579	0.13671	
☐	Model 2	Radial Basis Function	3.34231	0.4129	540	0.01517	0.1443	0.46605	0.57729	0.1443	
☑	Model 3	Radial Basis Function	2.6661	0.32179	487	0.01896	0.13165	0.51433	0.49287	0.13165	Largest RSquare
☐	Model 4	Radial Basis Function	0.78753	0.39698	555	0.06448	0.12658	0.47862	0.1162	0.12658	
☐	Model 5	Radial Basis Function	0.50468	0.48312	627	0.11125	0.14177	0.44302	0.03584	0.14177	
☐	Model 6	Radial Basis Function	0.30394	0.25499	470	0.14159	0.13924	0.4422	0.13495	0.13924	
☐	Model 7	Radial Basis Function	1.07701	0.30909	483	0.05563	0.11899	0.49528	0.19118	0.11899	
☑	Model 8	Radial Basis Function	1.61959	0.21987	426	0.05183	0.11392	0.4901	0.25364	0.11392	Smallest Misclassification Rate
☐	Model 9	Radial Basis Function	0.14094	0.0259	413	0.25284	0.22532	0.38964	0.00074	0.22532	
☐	Model 10	Radial Basis Function	1.28614	0.10527	379	0.10493	0.1443	0.46291	0.29554	0.1443	
☐	Model 11	Radial Basis Function	1.81867	0.04989	350	0.12137	0.1519	0.40534	0.36916	0.1519	
☐	Model 12	Radial Basis Function	2.25263	0.1456	385	0.06195	0.12405	0.46067	0.27913	0.12405	
☐	Model 13	Radial Basis Function	4.22111	0.27001	444	0.01643	0.13418	0.49299	0.5502	0.13418	
☐	Model 14	Radial Basis Function	4.90951	0.33724	491	0.01517	0.14177	0.46228	0.58606	0.14177	
☐	Model 15	Radial Basis Function	3.60894	0.35498	501	0.01517	0.14177	0.48394	0.5703	0.14177	
☐	Model 16	Radial Basis Function	4.06865	0.49015	594	0.01517	0.14177	0.41829	0.58389	0.14177	
☐	Model 17	Radial Basis Function	3.7914	0.21049	416	0.02149	0.12911	0.49587	0.44722	0.12911	
☐	Model 18	Radial Basis Function	4.67451	0.1757	403	0.02149	0.13165	0.47827	0.43926	0.13165	
☐	Model 19	Radial Basis Function	4.50697	0.01144	342	0.15424	0.18481	0.30633	0.50117	0.18481	
☐	Model 20	Radial Basis Function	3.01529	0.08409	361	0.08217	0.13924	0.45176	0.32986	0.13924	

References

Conservation International. “Ocean Pollution: 11 Facts You Need to Know.” https://www.conservation.org/stories/ocean-pollution-11-facts-you-need-to-know. Accessed 10. Oct. 2021.

CSIRO Oceans and Atmosphere, et al. An Analysis of Marine Debris in the Us. Oceans and Atmosphere, CSIRO, 2018.*WorldCat*, https://marinedebris.noaa.gov/reports/analysis-marine-debris-us. Accessed 10 Oct. 2021.

“[Data] Marine Debris Monitoring and Assessment Project (MDMAP) Accumulation Report: Plastic Pollution.” *Earth Challenge 2020 Working Version*, https://globalearthchallenge.earthday.org/datasets/EC2020::data-marine-debris-monitoring-and-assessment-project-mdmap-accumulation-report-plastic-pollution/about.Accessed 10. Oct. 2021.

EMIRHAN BULUT, “NASA Project; Marine Debris Machine Learning.” Kaggle, 2021, doi: https://www.kaggle.com/emirhanai/nasa-project-marine-debris-machine-learning. Accessed 10.Oct. 2021.

Mghili, Bilal, et al. “Marine Debris in Moroccan Mediterranean Beaches: An Assessment of Their Abundance, Composition and Sources.” *Marine Pollution Bulletin*, vol. 160, 2020, p. 111692., https://doi.org/10.1016/j.marpolbul.2020.111692. Accessed 10 Oct. 2021.

Program, NOS OR&R Marine Debris. “OR&R's Marine Debris Program.” NOAA *Marine Debris Program*, https://marinedebris.noaa.gov/. Accessed 10. Oct. 2021.

Rosevelt, C., et al. “Marine Debris in Central California: Quantifying Type and Abundance of Beach Litter in Monterey Bay, CA.” *Marine Pollution Bulletin*, vol. 71, no. 1-2, 2013, pp. 299–306., https://doi.org/10.1016/j.marpolbul.2013.01.015. Accessed 10 Oct. 2021.

Slatt, Roger M. “Barrier Island.” *Barrier Island - an Overview | ScienceDirect Topics*, Landscape Evolution in the United States, 2013, https://www.sciencedirect.com/topics/earth-and-planetary-sciences/barrier-island. Accessed 18. Nov. 2021

Van Truong, Nguyen, and Chu Beiping. “Plastic Marine Debris: Sources, Impacts and Management.” *International Journal of Environmental Studies*, vol. 76, no. 6, Dec. 2019, pp. 953–973. *Environment Complete*, EBSCOhost, https://doi.org/10.1080/00207233.2019.1662211. Accessed 10 Oct. 2021.