CLEANING UP COASTLINES

Brandon Bennitt
bbennitt@hawk.iit.edu

Yashwanth Praveen Pasupuleti
ypasupuleti@hawk.iit.edu

ABSTRACT

As humans continue to pollute the earth with plastic and other garbage, our oceans continue to get filled with more debris that does not belong there. This debris is harmful to ocean life and therefore the earth as a whole. In this project, we wish to provide a means of reducing the amount of plastic entering our oceans by estimating the amount of labor needed to clean a certain beach at a given time of year. Throughout this project, we try to prove relationships between non-traditional variables and the amount of garbage on the beach, such as the ruling political party in the state the beach is in. In the end, we are able to produce a model that can predict the amount of trash on a given beach that has an adjusted R2 value of .2319. Since our model did not have good predictive power, we decided to forgo estimating the amount of labor needed to clean the beaches to avoid producing misleading numbers. Instead, we spend time diagnosing our model and discussing how we can produce a model that has more predictive power in the future.

1 Introduction

Every year, an estimated 8 million metric tons of plastic end up in our oceans. Over $90 billion is being spent on cleaning up ocean trash, better managing waste, improving water treatment plants, and research to combat this problem. This sum can be drastically decreased in the future by reducing the amount of debris that enters the oceans. One of the many ways the amount of trash entering the ocean can be reduced is organizing clean up efforts on beaches and coastlines. By being able to predict the amount of trash at a given location, organizations can estimate the amount of people needed to clean that location before the garbage can enter the ocean. For this project, our main goal is to predict the amount of money needed to clean up a beach/coastline in a given city in the United States.

Besides just providing predictive power to clean up groups, we try to gain insight into why certain beaches have more trash than others. For this reason, we spend a lot of time on data exploration and will focus on building linear models for high explanatory power. Some of the analysis that we perform includes analyzing the effect of county population, ruling political party, and greater overall effect of the state the beach lies in. In addition, we try to show which counties and states have a high percentage of plastic in the garbage collected in previous clean up efforts. If we recognize a certain county or geographic area has a high percentage of plastic collected, we can recommend legislation be put in place to discourage the use of plastic.

2 Related Work

Since cleaning up pollution and plastics in the ocean is not a new concept, we first analyze the work of previous researchers who have conducted studies in this area.

Dr. Marcus Eriksen has been a guest lecturer at top universities such as MIT and has conducted a study that analyzes data from five years of observational work at sea where he collected data on pollution in the oceans. Dr. Eriksen's main goal is to place a number on the amount of plastic that is in the ocean as well as the percentage that is microplastics below a certain threshold. Our main take away from this research is that plastics that enter the ocean often degrade and become too small to

collect. Therefore, we see there is an opportunity to help the effort by collecting the plastic before it can even get into the ocean.

In a paper written by Colette Wabnitz from Stanford University, the problem of plastic pollution and its effects are further explored. The author specifically writes a section about removal of plastics from ocean fronts and beach clean up effectiveness. As Wabnitz is an expert in the field, we feel that if she is describing it as a potential way to help solve the problem, then it is worth contributing our time and efforts to a project that has the potential to make a difference.

In an article written by Hannah Ritchie, there are many visualizations that help the reader easily understand the problems, solutions, and improvements that can be made to help the plastic pollution problem. Reading this article made us realize the power of converting data and words into visualizations for people to understand the true meaning of what is being explained. Therefore, we draw inspiration from this article to make visualizations to show different relationships between variables.

In a paper written by Marcus Haward titled "Plastic pollution of the world's seas and oceans as a contemporary challenge in ocean governance", Haward discusses the issues that need to be addressed by legislators or organizations such as the International Maritime Organization (IMO). He discusses the current actions that have been taken to try and improve effectiveness and concludes by suggesting international agreements that can occur to help the issue. While we are not concerned with providing a model or analysis to influence how international agreements move forward, we draw inspiration from the idea that legislative action can be taken to help reduce plastic pollution. For this reason, we explore if there is a relationship between the ruling political party and overall average garbage collected in each state.

3 Data Processing

To begin our project, we researched for a dataset that had historical data on garbage found on beaches. We were able to find one relevant dataset that consists of rows that are individual records of clean up efforts on beaches around the world. This dataset was created from three other datasets from different international organizations around the world that collected this data in a joint effort and contained 55,000 rows. Luckily, the dataset was already cleaned by other researchers who used the data for their own needs. Unfortunately, we wanted to shrink the dataset to contain only entries from the U.S., which left us with only 2,000 rows.

In our remaining 2,000 rows, we had 16 columns for each row. The features that were in each row are shown in the table 1 below.

Table 1: Feature Descriptions

Name	Description
X	Specifies the location latitude
Y	Specifies the location longitude
SubCountry_L1_FromSource	The state the beach is in
SubCountry_L2_FromSource	The county the beach is in
TotalWidth_m	Total width of the beach in meters
TotalLength_m	Total length of the beach in meters
ShorelineName	Name of the specific beach
TotalVolunteers	Number of volunteers that helped with the recorded task
Year	Year which the cleanup took place
MonthNum	Number 1-12 to represent the month the cleanup took place
Day	Number 1-31 to represent the day of the month the cleanup took place
TotalItems_EventRecord	Total number of items that were collected as garbage
TotalClassifiedItems_EC2020	Total number of items that were classified into a category
PCT_PlasticAndFoam	Percent of items classified as plastic or foam
PCT_Glass_Rubber_Lumber_Metal	Percent of items classified as glass, rubber, lumber, or metal
LAND_TYPE	Type of land. Primary land, small island, large island
LAND_RANK	The land type encoded into numerical values

Before moving forward, we did some basic checks on our dataset to ensure we would not encounter any difficulties as we moved forward. After doing some visual and automated checks, we realized that it would be good practice to standardize the column names to lowercase values so our dataset would be easier to work with. After we did this, we converted columns to their respective data types. For example, we made sure to make categorical columns factors. Once that was complete, we went ahead and assumed that our data was ready for analysis.

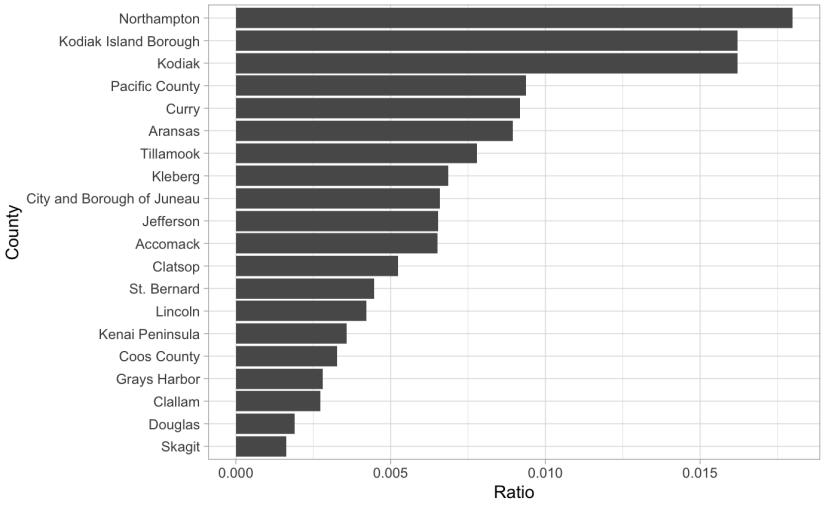
As discussed later in the paper, one of our goals of this project is to analyze the effect of county population and political party on garbage found on beaches. We also intend to give a cost estimate to an organization that pays employees minimum wage to clean up beaches. Therefore, we had to collect the missing data ourselves. We found a census dataset that included the population in every county in the U.S. and were able to join that dataset with our initial dataset to get the population of each county. To get the ruling political party, we had to create our own dataset that included the state name and the electoral college nominee for each state in the 2020 presidential election. This was a tedious task as we had to do some manual data entry. To get the minimum wage data, we performed the same type of manual data entry. We researched the minimum wage in every county that was present in our initial dataset and added the data to a new csv. These two manually generated csv files were then loaded into R as a dataframe and joined onto the initial dataset to get our combined dataset we continued to work with the rest of the project.

As we were joining the datasets, we noticed a few interesting details about some of the columns. We had multiple rows that were supposed to be representing the same state or county, but had slightly different notations. For example, we noticed that Hawaii was represented by 'HI', 'HA' and 'HW'. At this point we went back and cleaned the columns to set all similar states to their proper names.

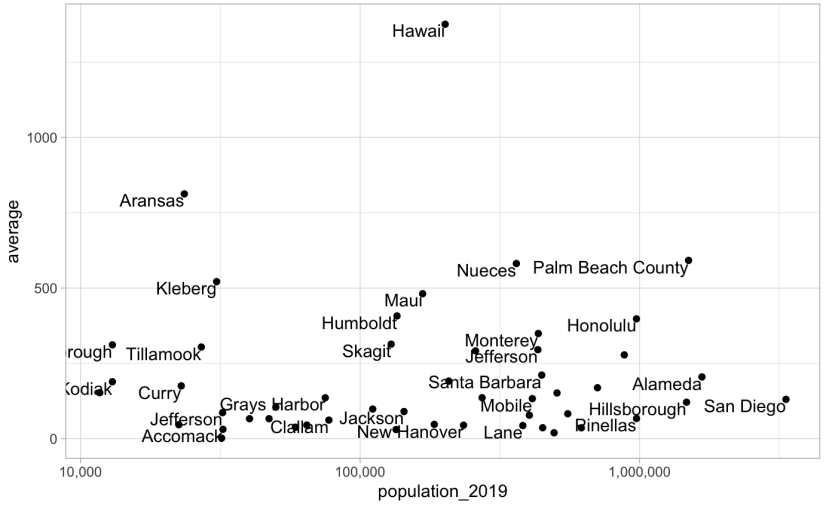
In terms of our data pipeline, we had stored a copy of the initial dataset and made sure to not alter the original csv in any way. We did the same with the other two csv files that we were using in our analysis. To ensure that the process of cleaning and joining the data together was reproducible, we kept the code at the beginning of the Rmd file where we did our data analysis and model building. Therefore, if at any point we wrote code that altered our dataframe we were working in, we would have a way to reset it and start from our cleaned, combined dataset instead of from square one.

4 Data Analysis

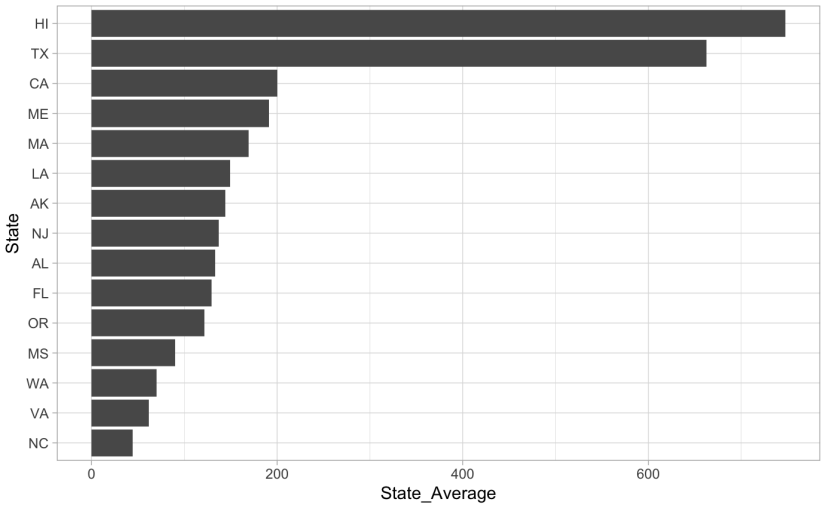
The first goal of our analysis will be to have a metric to rank the beaches based on the plastic that is being dumped on its shore, but going with the highest total number of plastic items dumped will be naive. This is because it also depends on the population, so we have decided to create a plastic dumped to population ratio for each county. This ratio is not very accurate, but can give the general sense of the plastic that's being dumped compared to the population. We assumed that the population of the county in which the beach is present is primarily responsible for the plastic dump on the shore. Below are the results with Northampton county coming in at number one based on the metric.



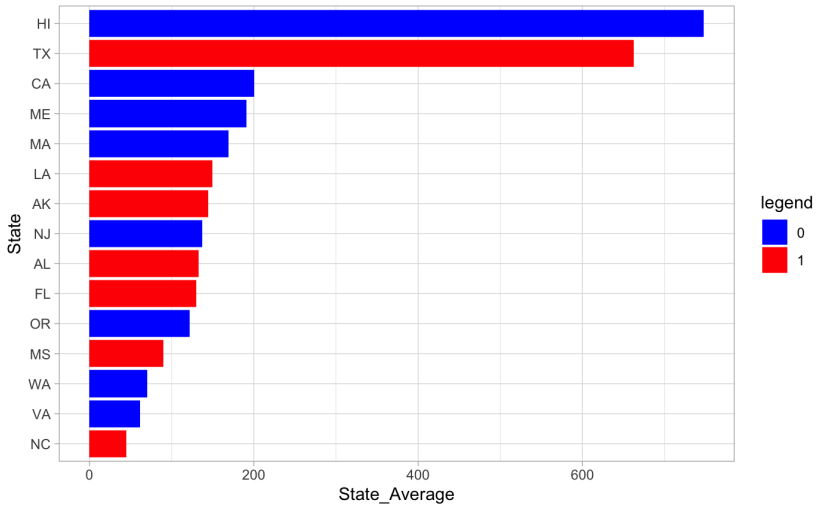
Our next goal was to look at the average garbage found on the beach with respect to the population of the county but not in the form of ratio. We wanted to use a scatter plot because then we would be able to visualize which counties with low population have the most garbage found on its shore. Below is the plot and Aransas county and Hawaii county are notable in this regard. Note that to fit more points into the plot, we used a logarithmic scale for population.



To further investigate, we wanted to look at the average plastic garbage collected per collection per state. This is meaningful because we were curious to see which state has the most garbage on its beaches' shores. Hawaii stood first and we were not shocked as almost the entire state is a beach.



Now that we have the plot showing the average garbage collected per collection in each state, we thought it would be interesting to see how these states voted in the previous election. We can see the results below where 0 is democratic and 1 is republican.



By merely looking at the plot there is not much that we could conclude in terms of correlation except for the fact that it is not extremely one sided. To quantify the relationship, we performed a hypothesis test to compare the averages of the garbage found on the beaches with the same political affiliation.

```
        Welch Two Sample t-test

data:  pol_data$average by pol_data$political_affiliation
t = 0.17001, df = 12.902, p-value = 0.8676
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -222.5926  260.5868
sample estimates:
mean in group 0 mean in group 1
       212.3593        193.3622
```

Looking at the p-value, which is not less than 0.05, we conclude that the amount of garbage found on the shore is independent of the party that received the state's electoral vote in the previous election. By saying so, we know that not one party is to blame over the other in terms of legislative action that may be put in place to reduce garbage on shorelines.

At this point, we were able to show the effects of population and political affiliation on the garbage collected on each beach. As a recap, we were able to show which counties had the highest garbage to population ratio as well as disprove that political affiliation is a direct factor in determining the garbage found on any one beach.

We also did some analysis on the percent of plastic found in each state. Our idea was to calculate the total amount of items collected in each state, the total number of plastic items collected in each state, and then divide the two to get a percentage. Unfortunately our data was never fit for this type of calculation as there were many rows which did not have classified items and they only included the total number of items they collected. Some columns that did classify certain items had several that were not classified. This made the task very challenging so we decided to forgo it all together. If we had the appropriate data, this would be a key metric we would look at to reccommend legislative action takes place in certain states that were above a threshold value for being a large plastic polluter.

5 Model Training and Model Diagnosis

5.1 Base Model

Before we trained our model, we performed a test/train split to ensure we can validate our performance on a testing set. We split the data as 75% for training and 25% for testing.

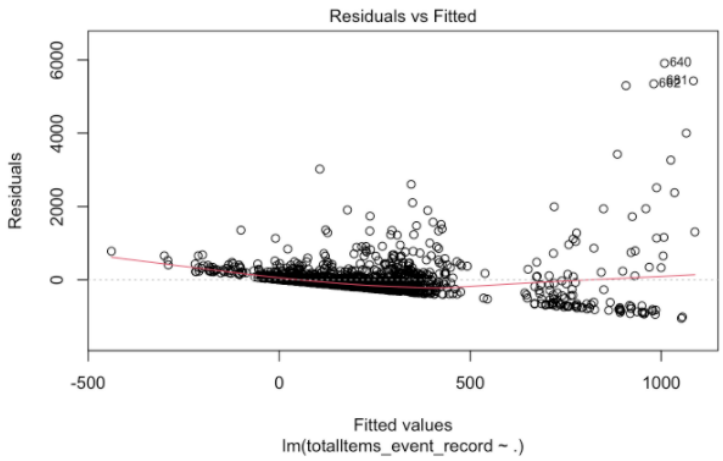
With first model that we built was our base linear model with the predictors as x, y, day of the week (whether if it's a weekday or weekend), time of the day, the county's population in which the beach is present, political party that's in power over the county, and month. Some of the features that we used as inputs into this model were engineered as they were not in the initial dataset. We converted each day of the way into a binary variable representing if it was a weekday or weekend, thinking that on the weekend we would tend to see a different pattern of garbage than on the weekdays. We converted the time of day into minutes from the start of the day to. Therefore, we would be able to know when our collection took place down to the minute of each day. The other variables were given in the initial dataset.

Below are the results from training the model.

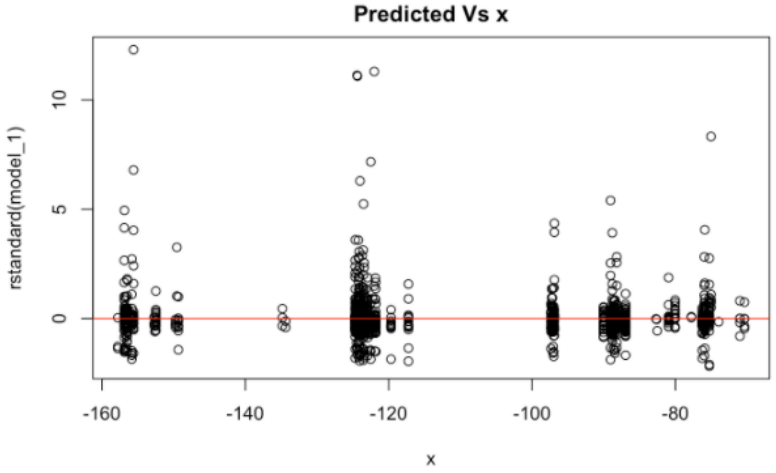
```
Residuals:
    Min      1Q  Median      3Q     Max
-1048.6  -177.9   -62.9    54.4  5907.6

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             3.892e+02  1.253e+02   3.106 0.001929 **
x                      -6.030e+00  7.021e-01  -8.588  < 2e-16 ***
y                      -2.534e+01  1.904e+00 -13.307  < 2e-16 ***
total_area_sq_m        -3.114e-04  1.245e-04  -2.500 0.012507 *
political_affiliation1  1.689e+02  4.985e+01   3.389 0.000719 ***
population_2019        -9.683e-05  3.690e-05  -2.624 0.008775 **
dow0                   -6.055e+01  2.596e+01  -2.333 0.019802 *
minute_of_the_day       4.891e-02  8.824e-02   0.554 0.579438
monthAug               -8.515e+01  5.819e+01  -1.463 0.143601
monthDec               -7.776e+01  6.346e+01  -1.225 0.220650
monthFeb               -1.159e+02  5.821e+01  -1.991 0.046607 *
monthJan               -3.635e+01  5.764e+01  -0.631 0.528364
monthJul               -5.701e+01  5.875e+01  -0.970 0.332002
monthJun                2.007e+01  5.774e+01   0.348 0.728253
monthMar                2.247e+01  5.756e+01   0.390 0.696338
monthMay               -4.605e+01  5.653e+01  -0.815 0.415402
monthNov               -8.415e+01  6.433e+01  -1.308 0.191080
monthOct               -1.059e+02  5.965e+01  -1.775 0.076063 .
monthSep               -7.624e+01  6.191e+01  -1.232 0.218317
land_rank3              1.579e+02  3.528e+02   0.447 0.654583
land_rank4              7.214e+01  8.345e+01   0.864 0.387462
land_rank5              2.078e+02  5.748e+01   3.615 0.000310 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 484.9 on 1517 degrees of freedom
  (21 observations deleted due to missingness)
Multiple R-squared:  0.1776,    Adjusted R-squared:  0.1662
F-statistic:  15.6 on 21 and 1517 DF,  p-value: < 2.2e-16
```

Straight away from the p-values we can conclude that most of the predictors are not as good as we thought they would be. Also, the R squared value is 17%(roughly speaking). This means only 17% of the variance of the predicted variable is being explained by the model. This means we have a bad model and need more features that are highly correlated with what we want to predict. One other solution may be we need more rows of data. This is also likely as we only have about 2,000 rows of data.

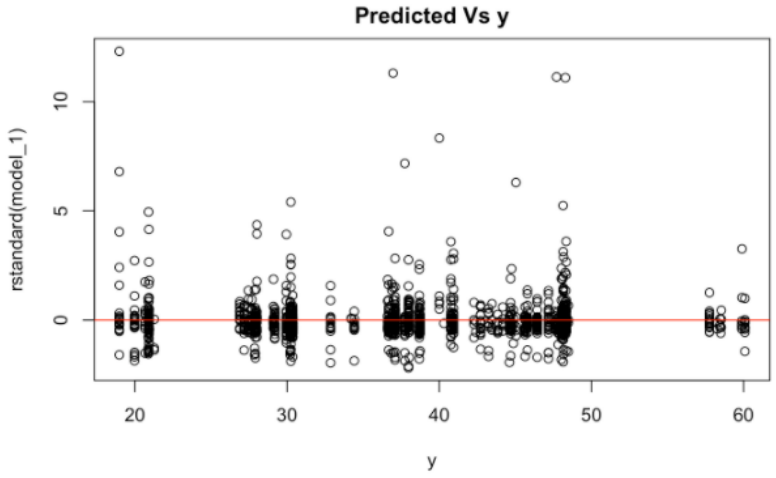
We can also take a look at the residuals below.



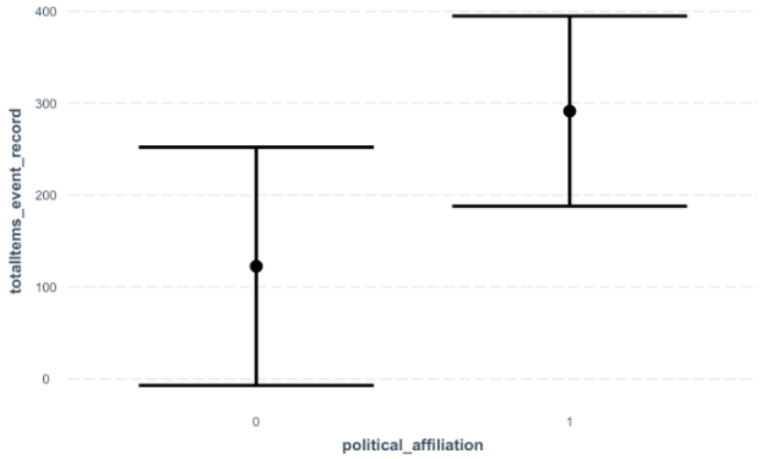
We can observe that there is not much of a pattern here. Since it is almost linear, we can establish that residuals are not absurd, but we need to look further to confirm.

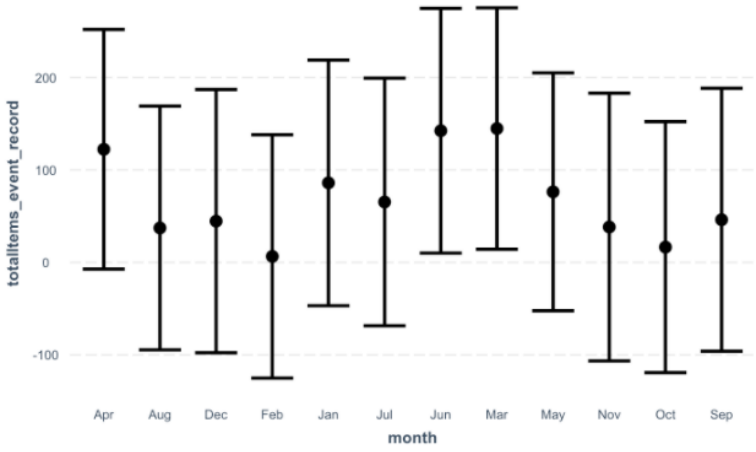


There also seems to be no pattern in the plot above. If we take a look at one more we can see the same thing.



To show how much the political affiliation and month was affecting the plot, we visualized the effect plots below.

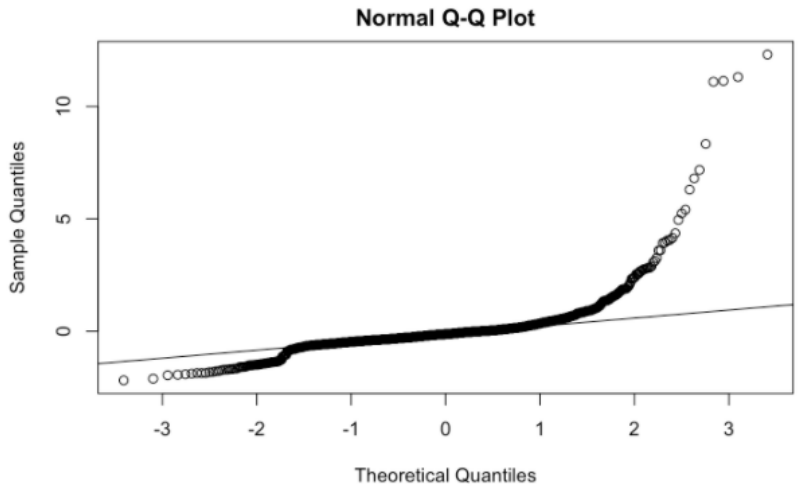




From the plots we can observe that the model is more likely to predict high number in total number of plastic items found in states where Republicans hold power as we can see that they have more positive effect on the model.

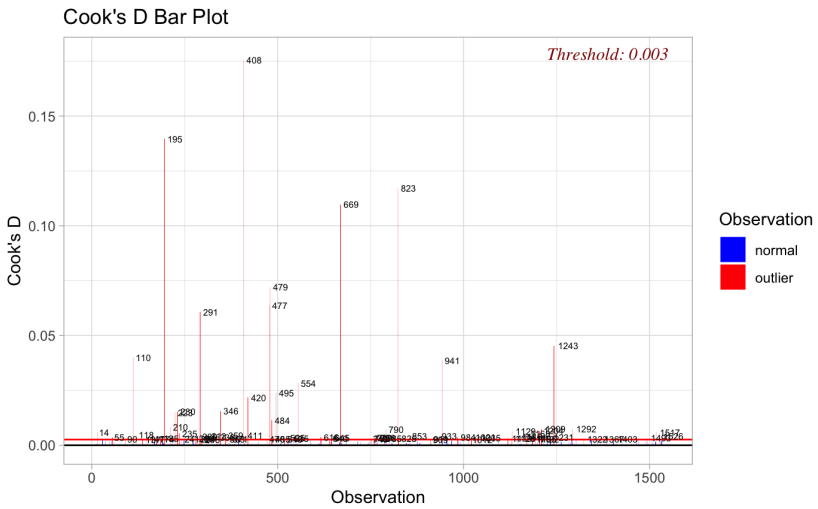
As one would assume, the months of March, May, and June positively affect the model indicating that the predicted value of total number of garbage collected could be higher than the rest of the months.

Let's look at the QQ plot to confirm the normality of the residuals.



The data points are pretty linear between -1 and 1, so it's safe to assume that the residuals are almost normally distributed.

Let's look at the influential points (we used cook's distance criteria to decide whether a data point is influential or not).



We can see that almost 10% to 15% of the data points are influential. This tells us that we need more data or it could also be that we need different data in order to predict.

Finally, we take a look at test RMSE values for this model.

```
> rmse_1<- sqrt( (y-y_1) %*% (y-y_1) / nrow(test) )
> rmse_1
         [,1]
[1,] 313.9229
```

As we can see here we have an RMSE value of 313.9229.

5.2 Transformed Model

To try and improve our results, we transformed our dependent variable, y, to (1/y). We built a stepwise model in order to see individual variable's effect on the model.

Below are the results.

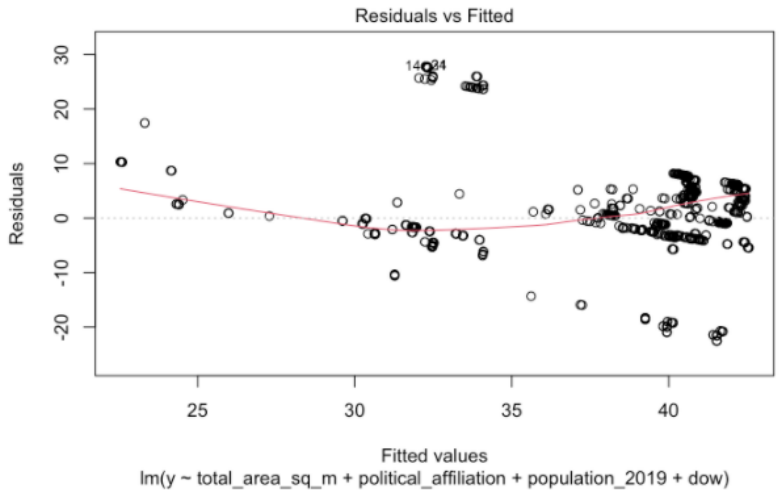
```
the response appeared on the right-hand side and was droppedproblem with term 3
in model.matrix: no columns are assigned
Call:
lm(formula = y ~ total_area_sq_m + political_affiliation + population_2019 +
    dow, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-22.5559  -3.1551  -0.8045   5.5884  27.8246

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             4.265e+01  3.732e-01 114.272  < 2e-16 ***
total_area_sq_m        -8.867e-06  2.003e-06  -4.426 1.03e-05 ***
political_affiliation1 -8.390e+00  4.591e-01 -18.273  < 2e-16 ***
population_2019        -5.520e-06  5.752e-07  -9.595  < 2e-16 ***
dow0                   -1.588e+00  4.313e-01  -3.682  0.00024 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.201 on 1555 degrees of freedom
Multiple R-squared:  0.2338,    Adjusted R-squared:  0.2319
F-statistic: 118.6 on 4 and 1555 DF,  p-value: < 2.2e-16

Start:  AIC=6976.22
y ~ total_area_sq_m
```

```
the response appeared on the right-hand side and was droppedproblem with term 3
in model.matrix: no columns are assigned                        Df Sum of Sq
RSS    AIC
+ political_affiliation  1   24306.5 111885 6671.5
+ population_2019        1    6491.5 129700 6902.0
+ x                      1    6316.1 129875 6904.1
+ dow                    1    2989.7 133202 6943.6
<none>                                136191 6976.2

Step:  AIC=6671.54
y ~ total_area_sq_m + political_affiliation

the response appeared on the right-hand side and was droppedproblem with term 4
in model.matrix: no columns are assigned                    Df Sum of Sq    RSS
AIC
+ population_2019  1    6389.0 105496 6581.8
+ dow              1    1108.3 110776 6658.0
<none>                         111885 6671.5
+ x                1       3.0 111882 6673.5

Step:  AIC=6581.82
y ~ total_area_sq_m + political_affiliation + population_2019
```

```
the response appeared on the right-hand side and was droppedproblem with term 5
in model.matrix: no columns are assigned       Df Sum of Sq    RSS    AIC
+ dow   1    911.69 104584 6570.3
<none>              105496 6581.8
+ x     1     86.37 105409 6582.5

Step:  AIC=6570.28
y ~ total_area_sq_m + political_affiliation + population_2019 +
    dow

the response appeared on the right-hand side and was droppedproblem with term 6
in model.matrix: no columns are assigned       Df Sum of Sq    RSS    AIC
<none>              104584 6570.3
+ x     1    96.377 104488 6570.8

Call:
lm(formula = y ~ total_area_sq_m + political_affiliation + population_2019 +
    dow, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-22.5559  -3.1551  -0.8045   5.5884  27.8246

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             4.265e+01  3.732e-01 114.272  < 2e-16 ***
total_area_sq_m        -8.867e-06  2.003e-06  -4.426 1.03e-05 ***
political_affiliation1 -8.390e+00  4.591e-01 -18.273  < 2e-16 ***
population_2019        -5.520e-06  5.752e-07  -9.595  < 2e-16 ***
dow0                   -1.588e+00  4.313e-01  -3.682  0.00024 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.201 on 1555 degrees of freedom
Multiple R-squared:  0.2338,    Adjusted R-squared:  0.2319
F-statistic: 118.6 on 4 and 1555 DF,  p-value: < 2.2e-16
```

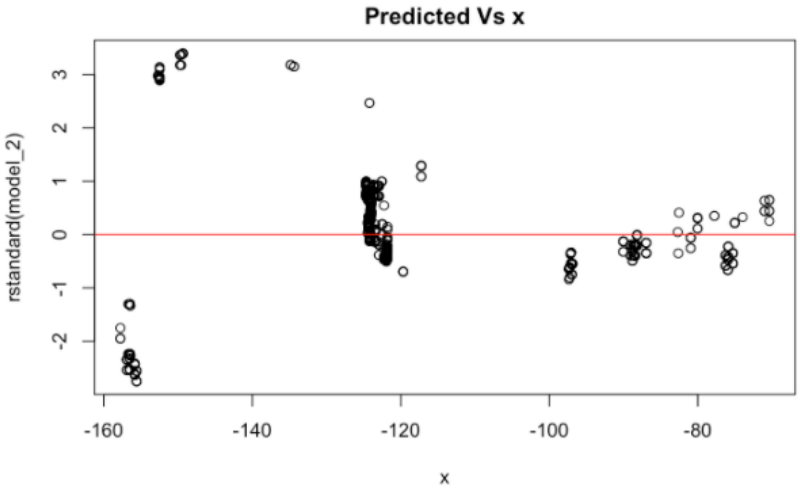
If we look at the AIC values, they Increase as the variables are being added, which is not good. The R squared value has now increased to 23.3%, where it was previously 17%. It is safe to say we also found the good predictors at the last iteration.

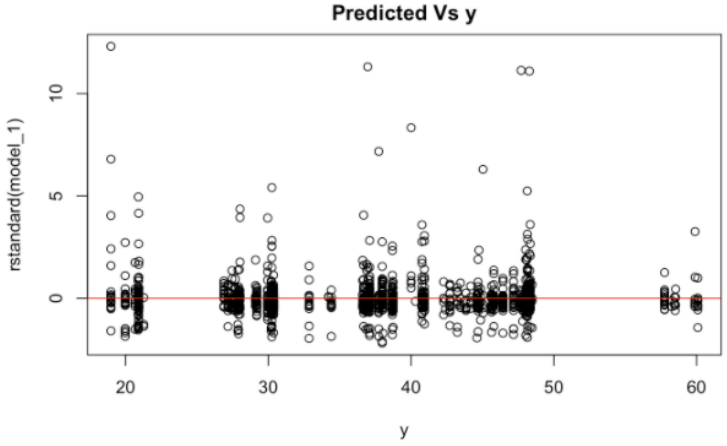
Let’s look at the model analysis.



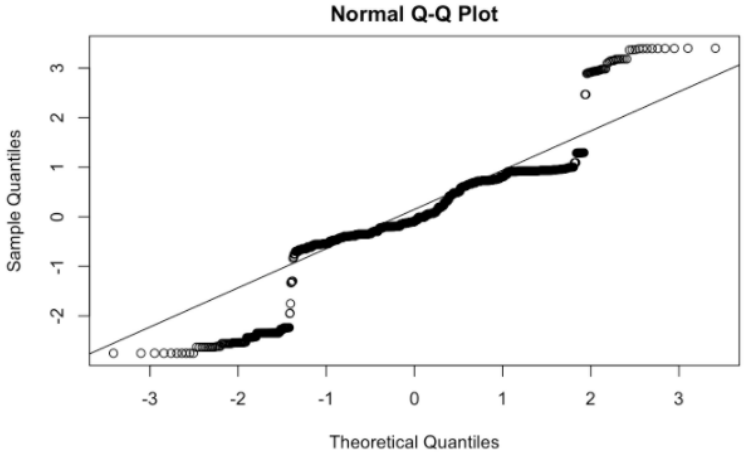
We see a slight U-shaped pattern but optimistically speaking we can say it is linear.

Looking at the predicted Vs X and Y plots:



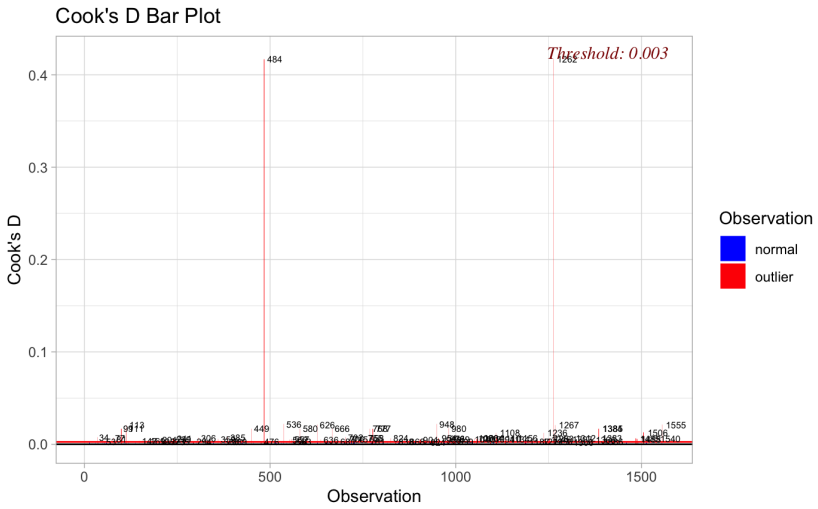


Looking at the QQ plot:



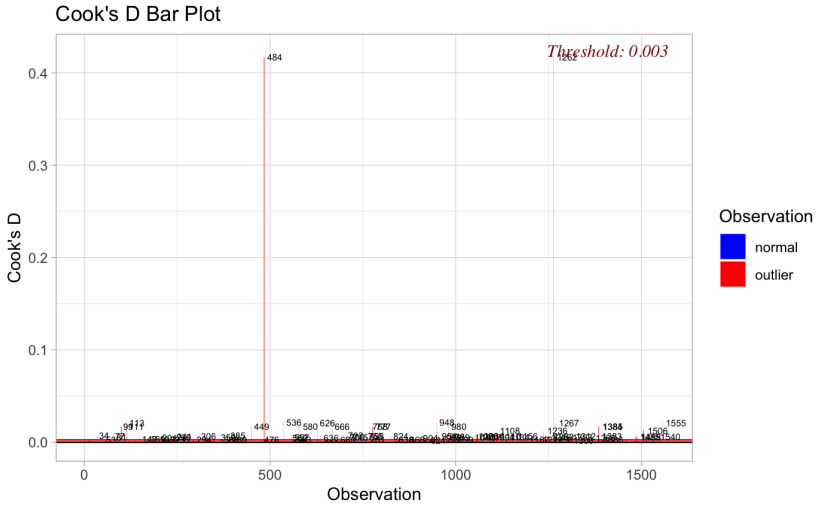
We can observe that the residuals are not so uniformly distributed at the tail ends of the distribution, but roughly 80% of the residuals are normally distributed.

Let's look at the influential points with the cook's distance plot.



We can observe that there are less influential points in here than that of the previous model for the same threshold.

Let's look at the test RMSE value for this model:



We can see here that we have a test RMSE of 362.5844. This is slightly higher than the previous model but is still comparable in terms of error.

6 Conclusion

As we can see in the results from the model training and analysis that we have done, we were not able to successfully build a highly predictive model with the variables that we had. We created two variations of a linear model to predict the amount of garbage we would find on a beach for a future clean up effort. With limited predicted power in our data, we focused our efforts more on inferential data analysis and increasing model accuracy with the data that we did have. This was our reasoning for using automated feature selection in the transformed model. We created multiple visualizations during our exploratory data analysis to show relationships, or lack of relationships, between different variables that we thought would be correlated. We explored how political affiliation does not have a direct correlation with garbage collected on a beach as proven by a hypothesis test.

In the end, we decided not to create a model for estimating labor required. Without large predictive power in our models, a labor estimation would prove to be useless and misleading to organizations that are looking for help.

To extend this project and for a chance at a higher predictive model, we can create a model with the entire dataset that contained global data. This would reduce sparsity in our dataset as we would have 55,000 rows instead of 2,000 rows. With more data, we expect to be achieve a much higher accuracy. Another way we can extend this project is by creating more features. We would have to obtain more data, but some features we thought of that might include more predictive power are days since last festival/holiday, weather data over the past 3 days, the number of garbage cans on the beach, and the types of plastics collected. Knowing the days since the last festival or holiday would allow us to see when the last time was that there were a larger number of people on the beach than normal. If we try and clean up the day after a festival, we are more likely to see more garbage. Weather data over the previous three days would allow us to see if there were an average number of people at the beach or above/below average. If it rained, we expect to see less garbage on the beach than if it was sunny. We believe if we can get the number of garbage cans on a given beach, we can show that people are more likely to throw their trash away if they see the opportunity to do so.

Lastly, it would be nice to create an estimate for the number of people needed to clean up a given beach.

All in all, we learned from our mistakes in this project and hope to extend our work in the future.

7 Data Sources

The initial dataset was originally found at https://globalearthchallenge.earthday.org/datasets/EC2020::data-earth-challenge-integrated-data-plastic-pollution-mlw-mdmap-icc-2015-2018/about.

The census dataset was found at https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html.

The other data was manually entered into a new csv from information found on various websites. This dataset is also stored in the project repository for reference.

8 Source Code

The source code along with the knitted .Rmd file for this project can be found at https://github.com/bbennitt/Cleaning-Up-Coastlines.

References

[1] Ocean pollution and marine debris. Ocean pollution and marine debris | National Oceanic and Atmospheric Administration. (n.d.). Retrieved December 6, 2021, from https://www.noaa.gov/education/resource-collections/ocean-coasts/ocean-pollution.

[2] Eriksen M, Lebreton LCM, Carson HS, Thiel M, Moore CJ, Borerro JC, et al. (2014) Plastic Pollution in the World's Oceans: More than 5 Trillion Plastic Pieces Weighing over 250,000 Tons Afloat at Sea. PLoS ONE 9(12): e111913. https://doi.org/10.1371/journal.pone.0111913

[3] Hannah Ritchie and Max Roser (2018) - "Plastic Pollution". Published online at OurWorldIn-Data.org. Retrieved from: 'https://ourworldindata.org/plastic-pollution' [Online Resource]

[4] Wabnitz, C. (2015). (PDF) editorial: Plastic pollution: An ocean emergency. ResearchGate. Retrieved December 6, 2021, from https://www.researchgate.net/publication/268187066_Editorial_Plastic_Pollution_An_Ocean_Emergency.

[5] Haward, Marcus. "Plastic Pollution of the World's Seas and Oceans as a Contemporary Challenge in Ocean Governance." Nature Communications 9, no. 1 (2018). https://doi.org/10.1038/s41467-018-03104-3.

[6] [Data] Earth Challenge Integrated Data: Plastic Pollution (MLW, MDMAP, ICC) 2015-2018. Earth Challenge 2020 Working Version. (n.d.). Retrieved December 6, 2021, from https://globalearthchallenge.earthday.org/datasets/EC2020::data-earth-challenge-integrated-data-plastic-pollution-mlw-mdmap-icc-2015-2018/about.