Cleaning Up Coastlines

By Brandon Bennitt and Yashwanth Praveen Pasupuleti

Presentation Outline

1. Introduction and Purpose
2. Proposed Solution
3. Project Timeline
4. Detailed View of Dataset
5. Preprocessing
6. Exploratory Data Analysis
7. Modeling and Model Results
8. Model Diagnosis
9. Conclusion and Future Works

Overview

Context

Every year, 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.

Problem statement

One of the many ways the amount of trash entering the ocean can be reduced is organizing clean up efforts on beaches and coastlines.

Organizations that help with these efforts are tasked with planning and assigning workers to multiple beaches. Often it is a challenge to know how many workers to assign to a beach.

Our Solution - Main Goal

Thought Process **Our**
Implementation to Solve Problem

- There are no/limited organizations that employ people to clean beaches
- With limited amount of people willing to volunteer we must maximize labor
- If we know how much garbage is on a beach and how large the beach is, we can estimate the amount of labor needed to clean the beach

- Use data from previous clean up events to predict how much garbage would be at a future event
- Given garbage that will be at each event, we can distribute volunteer labor to maximize our efforts
- With labor efforts as efficient as possible, we can help solve the pollution problem more effectively

Our Solution - Sub Goals

Questions **How to Answer**

1. What county has the highest garbage to population ratio?
2. Is there a relationship between the amount of garbage on a beach and the state that the beach is in?
3. Is the amount of garbage on the shorelines related to the political party that was nominated in the previous presidential election?
4. In certain states is it more likely to find a higher percentage of plastic in the garbage collected?

1. Compute average garbage collected per cleanup per county. Divide average garbage by county population. Create visualization
2. Create visualization to show average garbage per cleanup in each state
3. Use visualization from question 2 but color each state as political party. For quantitative results, run hypothesis test
4. Create visualization showing portion of total garbage collected that was plastic

Project Overview and Planning

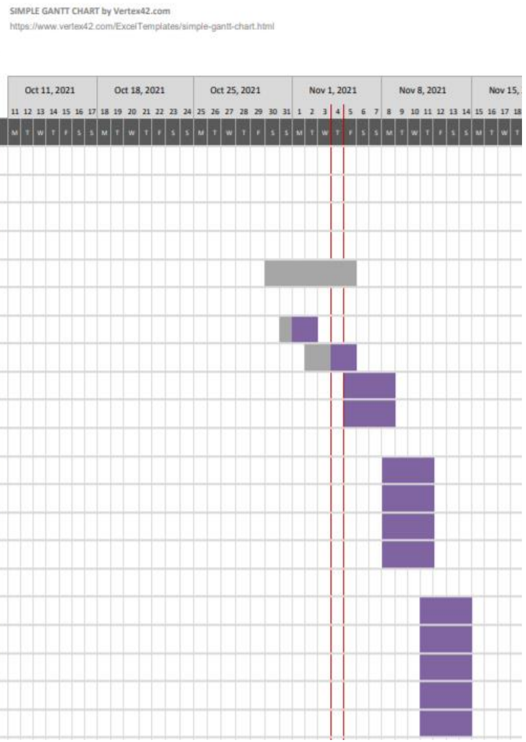
Initial Project Planning

CSP 571 Project Outline

Brandon Bennitt
Yashwanth Praveen

Project Start: Sat, 9/25/2021
Display Week: 4

TASK	ASSIGNED TO	PROGRESS	START	END
Preliminary Planning				
Select Group/Topic	Brandon and Yashwanth	100%	9/25/21	9/26/21
Data Gathering	Brandon and Yashwanth	100%	10/6/21	10/7/21
Project Outline	Brandon and Yashwanth	100%	10/8/21	10/10/21
Project Plan/Detail	Brandon and Yashwanth	100%	10/30/21	11/5/21
Data Cleaning				
Rename columns	Yashwanth	50%	10/31/21	11/2/21
Filter dataset to reduce redundant features	Yashwanth	50%	11/2/21	11/5/21
Adjust data types to accurately perform exploration	Brandon		11/5/21	11/8/21
Data Value Checking	Brandon		11/5/21	11/8/21
Data Exploration				
Calculate Summary Statistics	Brandon and Yashwanth		11/8/21	11/11/21
Discover distribution of each feature	Brandon and Yashwanth		11/8/21	11/11/21
Create simple plots to summarize data	Brandon and Yashwanth		11/8/21	11/11/21
Normalizing features	Brandon and Yashwanth		11/8/21	11/11/21
Data Visualization				
Create visualization for total garbage collected in each state	Brandon and Yashwanth		11/11/21	11/14/21
Create visualization for garbage per population in state	Brandon and Yashwanth		11/11/21	11/14/21
Create visualization for types of garbage collected in each state	Brandon and Yashwanth		11/11/21	11/14/21
Create visualization for political party in power vs amount of garbage	Brandon and Yashwanth		11/11/21	11/14/21
Create visualization to analyze relationship between neighboring states	Brandon and Yashwanth		11/11/21	11/14/21



Task Headings

- Preliminary Planning
- Data Cleaning
- Data Exploration
- Data Visualization
- Data Analysis
- Model Building
- Model Diagnosis
- Final Project Wrap Up

Challenges/Changes in Project Planning

Challenges **Solutions**

- Data cleaning and exploration took longer than anticipated
- Certain features had to be engineered starting with manual data collection
- Other projects and deadlines approached quickly
- Data did not contain as much explanatory power as we hoped

- Rescale our milestone deadlines to ensure sufficient time to clean and explore data
- Group effort to split monotonous work in half
- Communicate with team members about rescheduling meetings to work together
- Discuss what important features we may be missing and how to improve

Dataset Overview

Initial Dataset

Overview

- Global plastics dataset from 2015-2018
- Approximately 55,000 rows with entries from multiple countries around the world and across the entire US
- Only ~2,000 rows were entries from U.S. events
- U.S. data would allow for more inferential work

Initial Dataset

Columns

1. **(X,Y)** - Specifies the location coordinates
2. **SubCountry_L1_FromSource** (State)
3. **SubCountry_L2_FromSource** (City)
4. **TotalWidth_m** - width of the beach
5. **TotalLength_m** - length of the beach
6. **ShorelineName** - name of the beach
7. **TotalVolunteers** - Number of volunteers that helped with the recorded task
8. **Year** - year which the cleanup took place
9. **MonthNum** - number between 1-12 to represent the month the cleanup took place
10. **Day** - number between 1-31 to record the day of the month the cleanup took place
11. **TotalItems_EventRecord** - Total number of items that were collected as garbage
12. **TotalClassifiedItems_EC2020** - Total number of items that were classified into a category
13. **PCT_PlasticAndFoam** - Percent of items classified as plastic or foam
14. **PCT_Glass_Rubber_Lumber_Metal** - Percent of items classified as glass, rubber, lumber, or metal
15. **LAND_TYPE** - Type of land. Primary land

Manual Data Additions

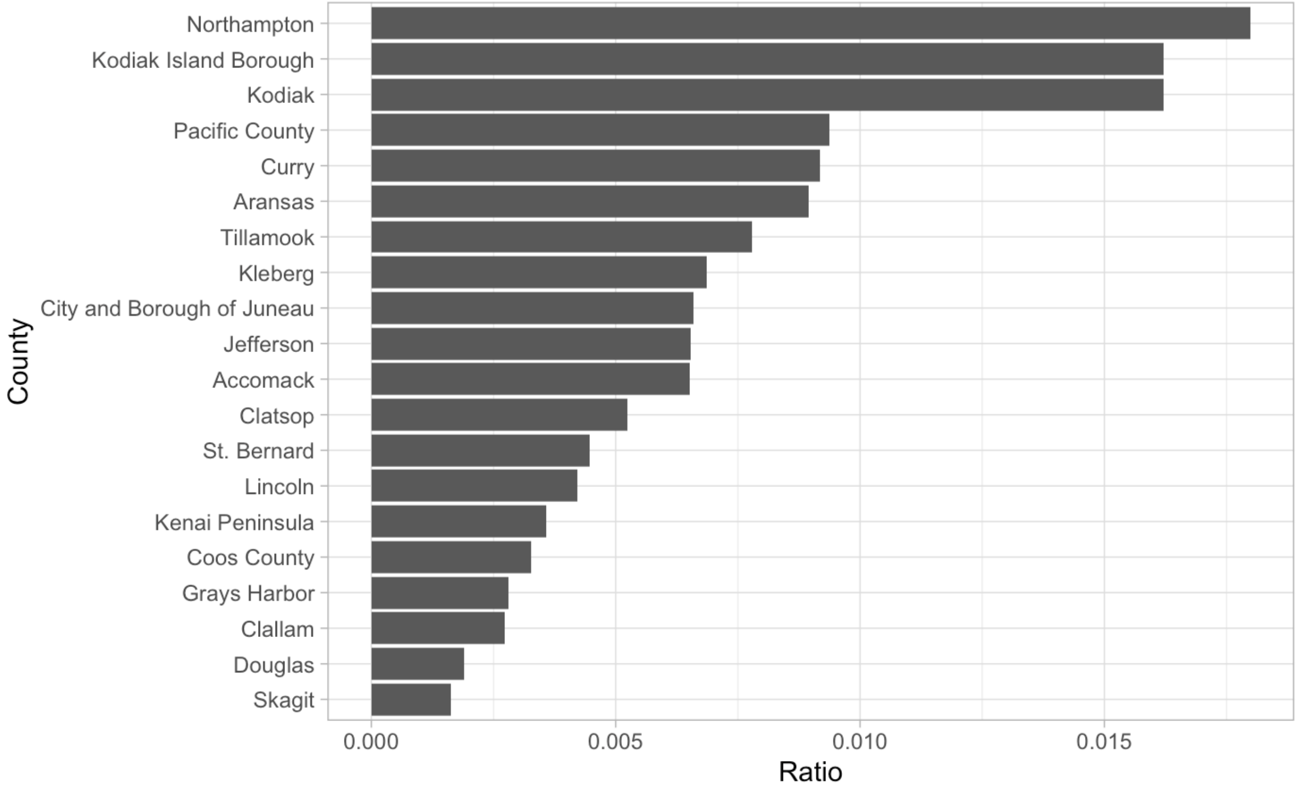
- Most sub-questions required additional data that was not in original dataset
    - Needed feature for political affiliation
    - Needed feature for population in each county
    - Needed minimum wage in each county to predict labor cost required
- Researched to find datasets
    - Found a census dataset containing the population of each county in the U.S. as of 2019
        - Write code to join column for population
    - Found dataframe containing 2020 election results for each state
        - Write code to join column for political affiliation
    - No dataset for minimum wage per county
        - Manually research and create csv with county and corresponding minimum wage

Data Cleaning and Preprocessing

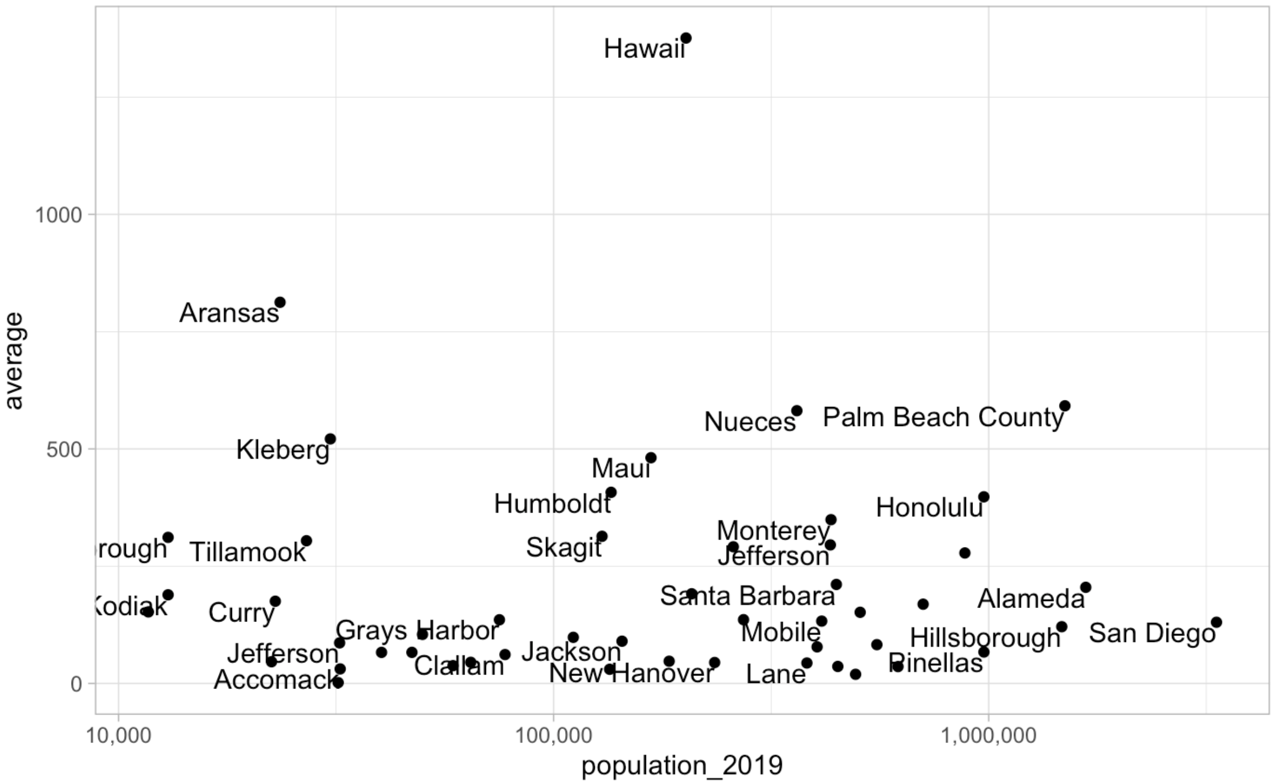
1. Cleaning and standardizing column names
2. Remove columns with N/A
3. Converting variables to appropriate data types
4. Filtering data pertaining only to U.S.
5. Unifying the state and county names
6. Merge data from different csv files to get required dataframe
7. Convert required character variables into factor variables

Exploratory Data Analysis

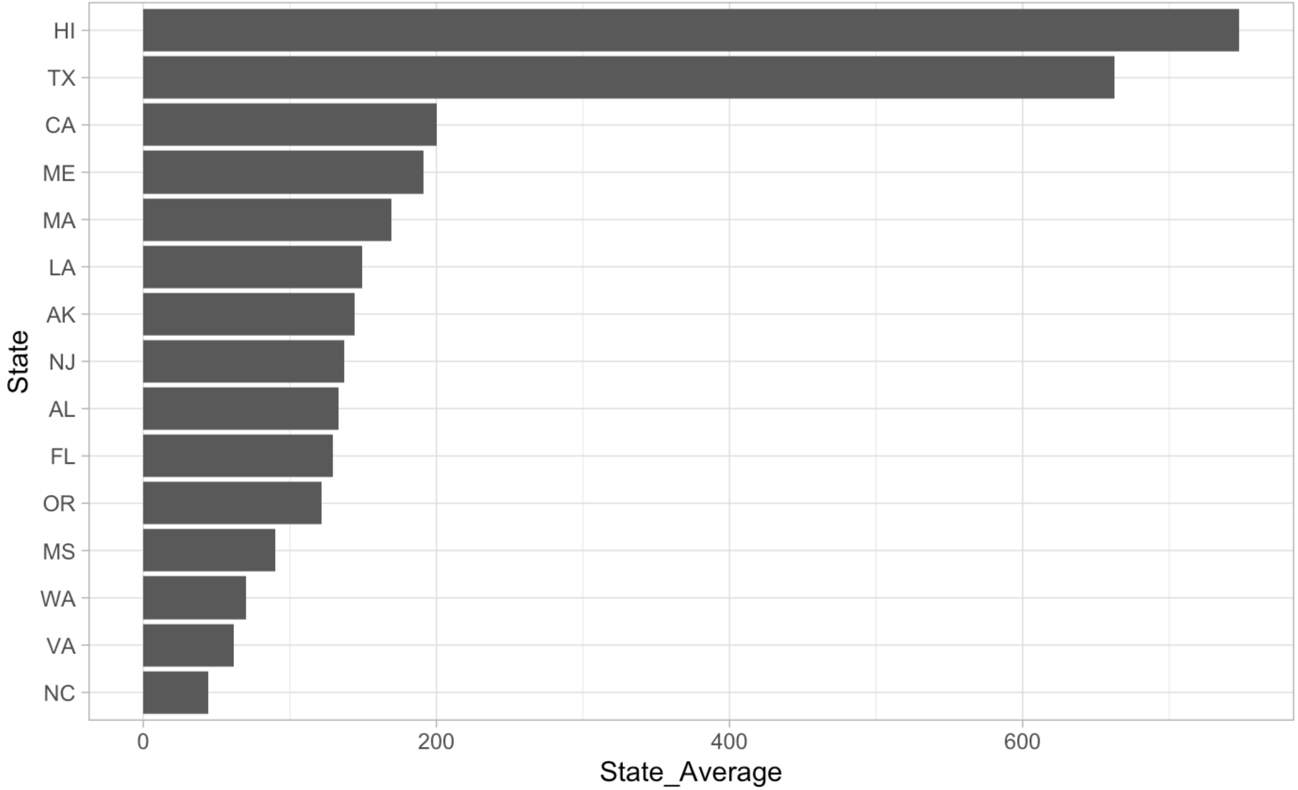
County Population to Garbage Ratio



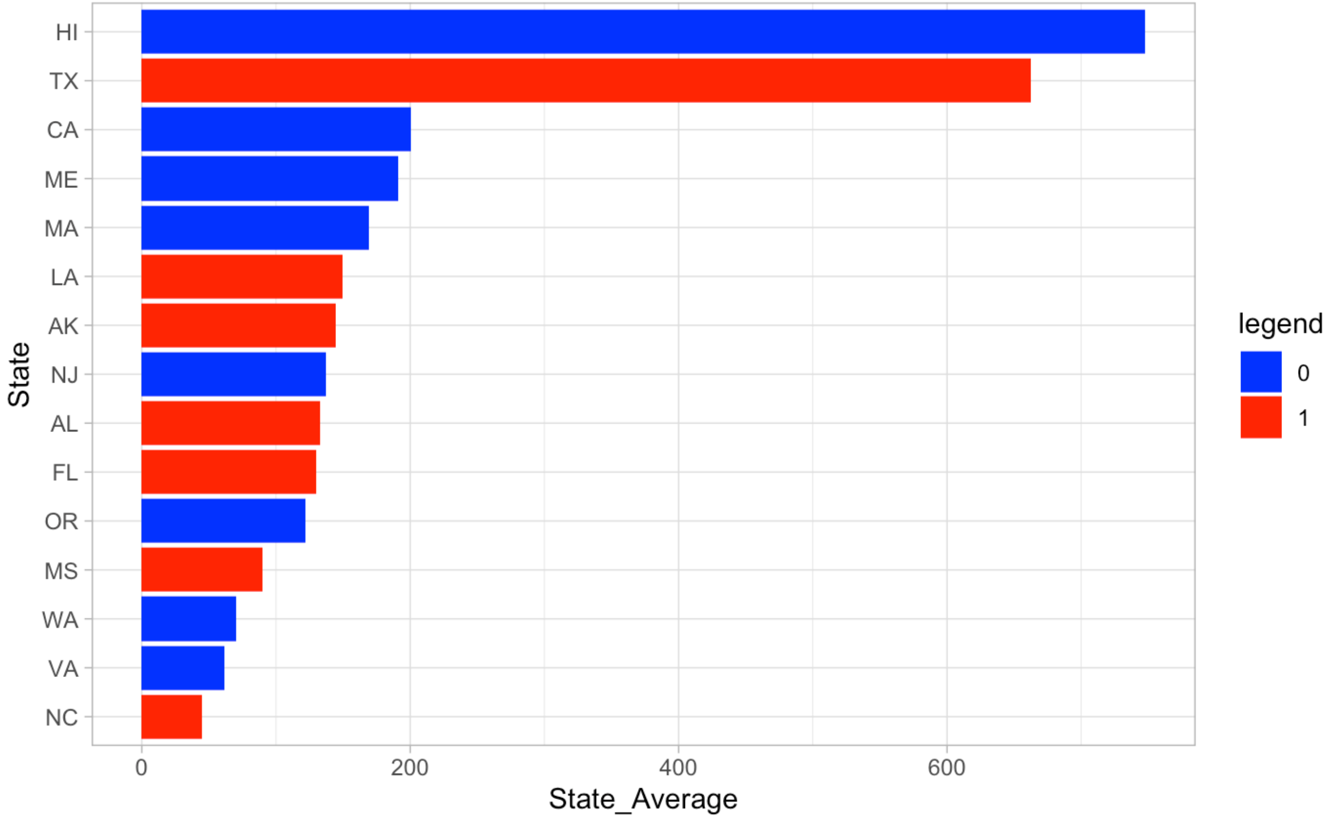
Relation Between County Population and Garbage Collected



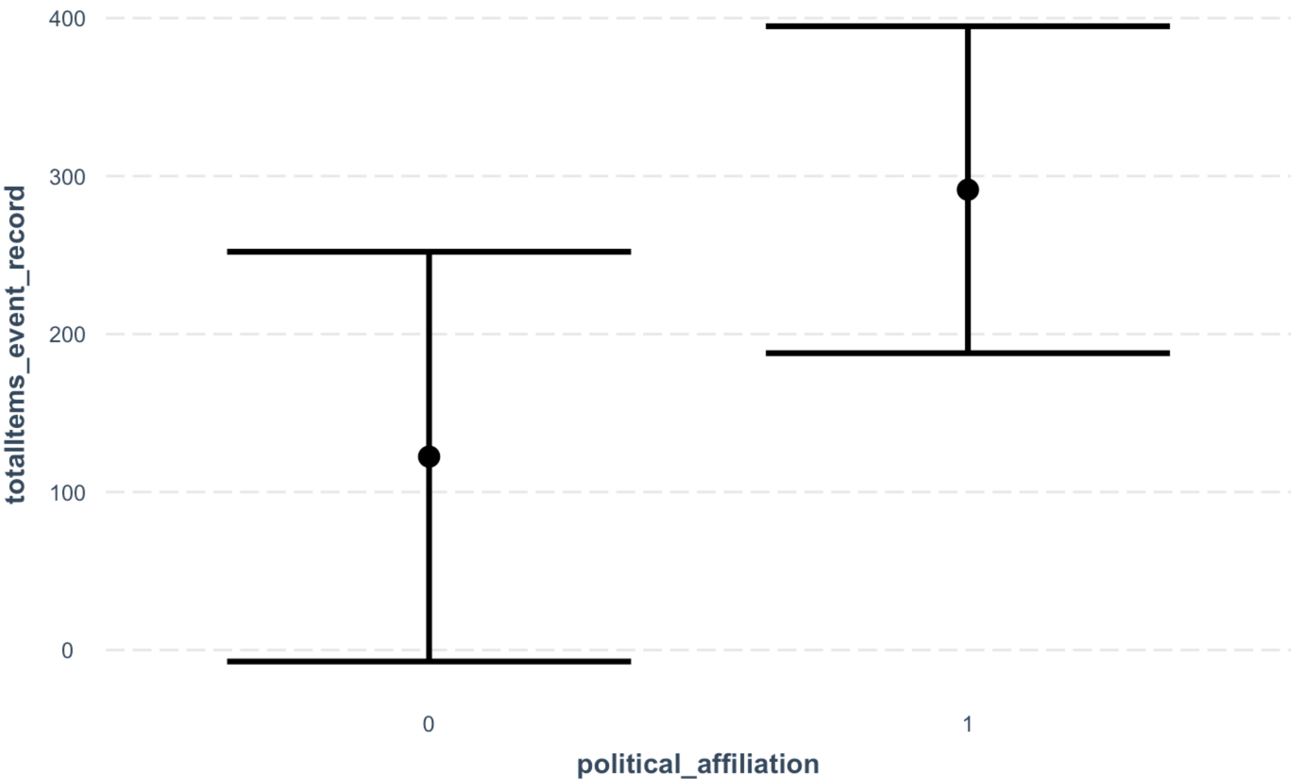
Average Garbage per Collection per State



Assigning Political Affiliation



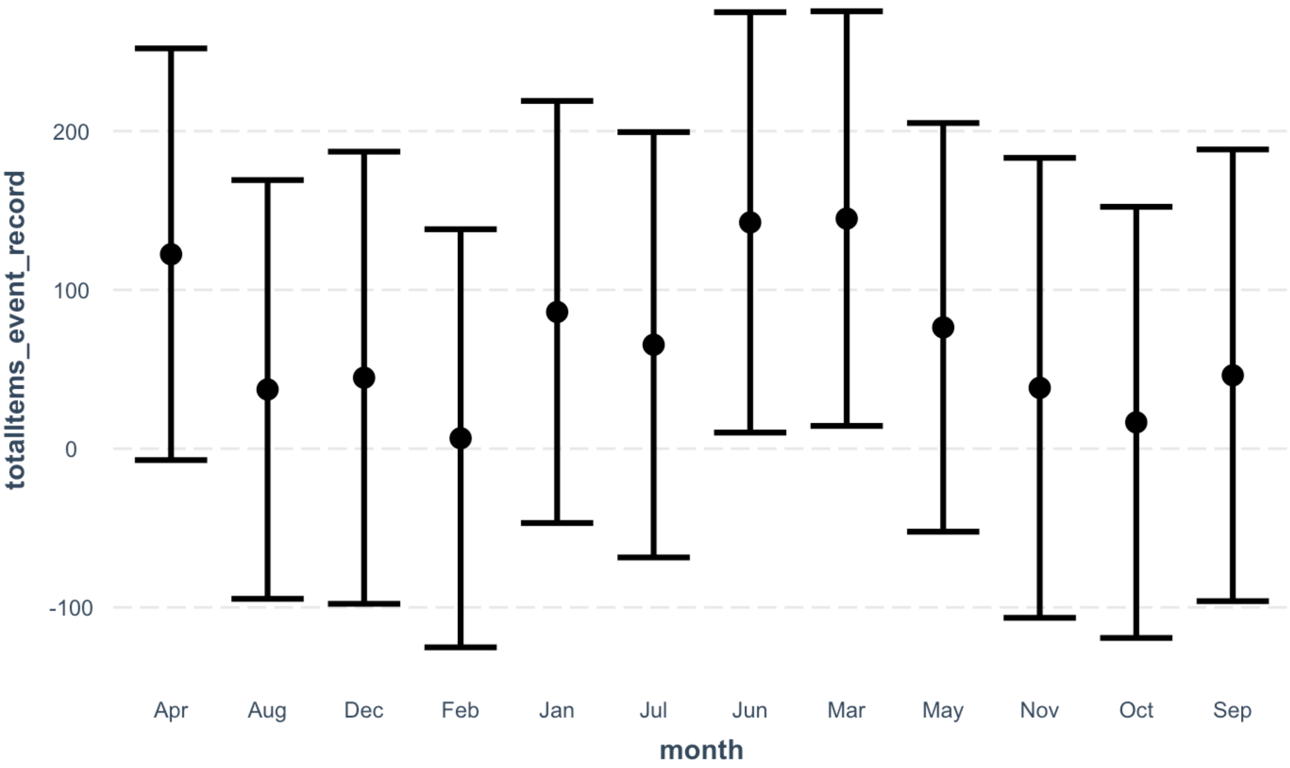
Effect of Political Affiliation on Garbage Collec



Hypothesis Test on 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 
```

Effect of Month on Garbage Collected



Percent of Plastic per State

- Plastic is largest material found in coastline debris
- Identifying which areas have a lot of plastic can lead to legislative action
  - No plastic straws in certain areas, etc.
- Tried to create bar chart visualization
- Unsuccessful since most rows did not contain appropriate classification of garbage collected
  - Saw 100% of plastic in a lot of cases, so not meaningful to compare

Modeling and Model Results

Base Model

- Y: total number of items collected on beach
- X_i: (latitude, longitude, minute of day, county population, political affiliation, total area of the beach, land rank)
    - Correlation plot showed little correlation between other features
- Linear model for high explanatory power

Base Model Results

```
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
```

Base Model Results

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

Transformed Model

- Y: 1 / total number of items collected on beach
- X_i: (latitude, longitude, minute of day, county population, political affiliation, total area of the beach, land rank)
  - Correlation plot showed little correlation between other features
- Linear model for high explanatory power

Transformed Model 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
```

Transformed Model Results

```
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
```

Transformed Model Results

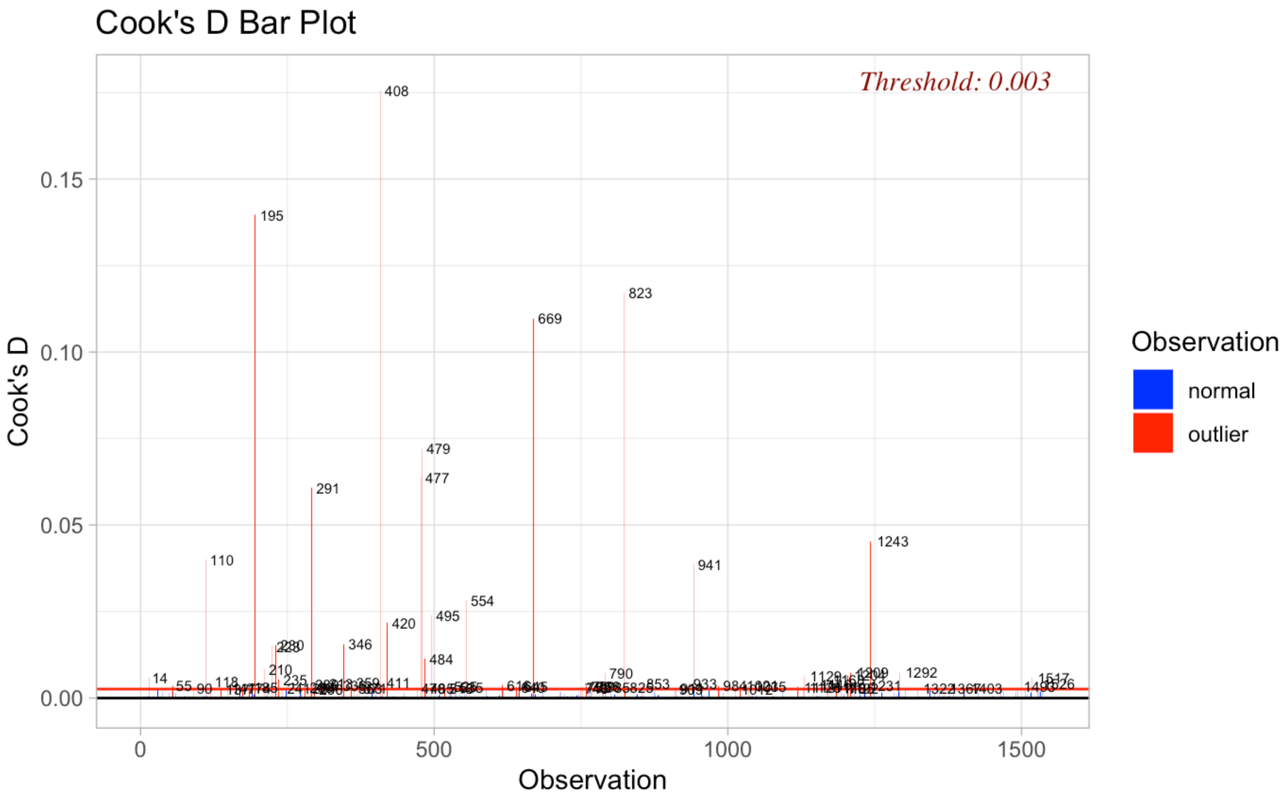
```
the response appeared on the right-hand side and was droppedproblem with term 5
in model.matrix: no columns are assigned       Df Sum of Sq    RSS    AIC
+ dow    1    911.69 104584 6570.3
<none>                105496 6581.8
+ x      1     86.37 105409 6582.5

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
```

Transformed Model Results

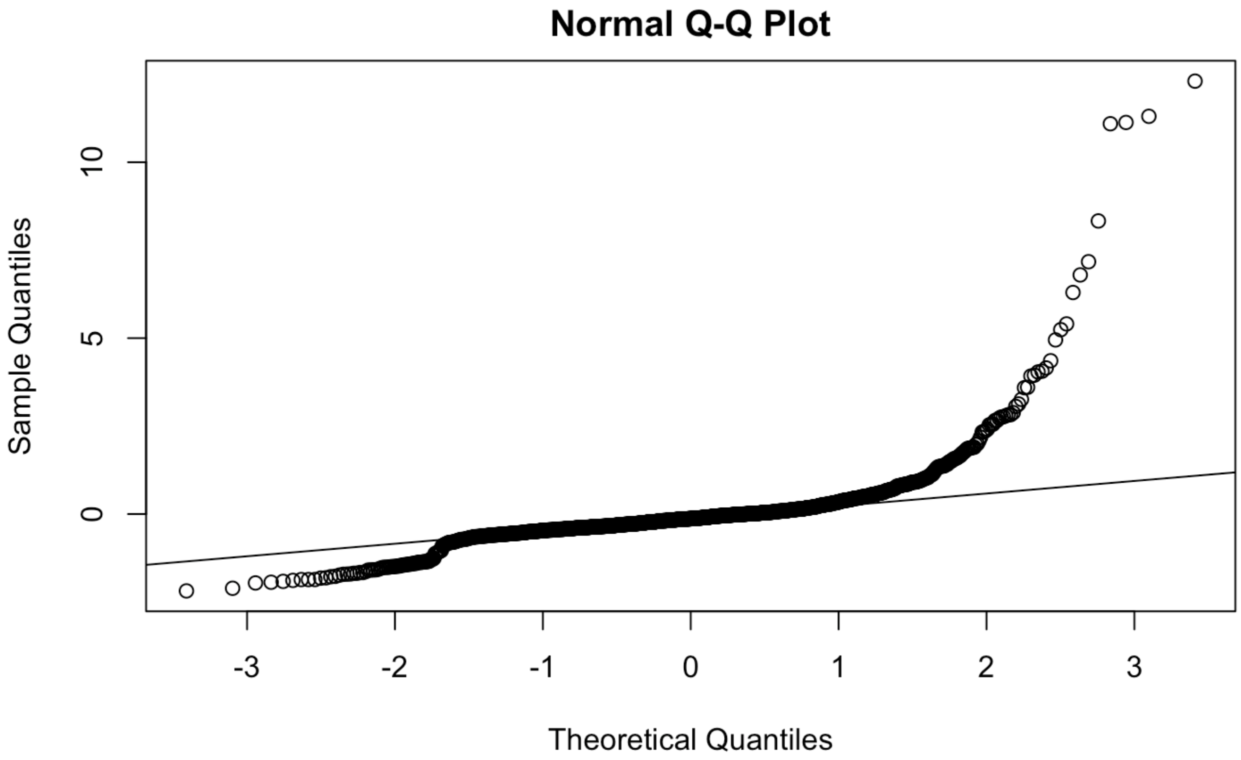
```
> rmse_2<- sqrt( (y-y_1) %*% (y-y_1) / nrow(test) )
> rmse_2
         [,1]
[1,] 362.5844
```

Model Diagnosis

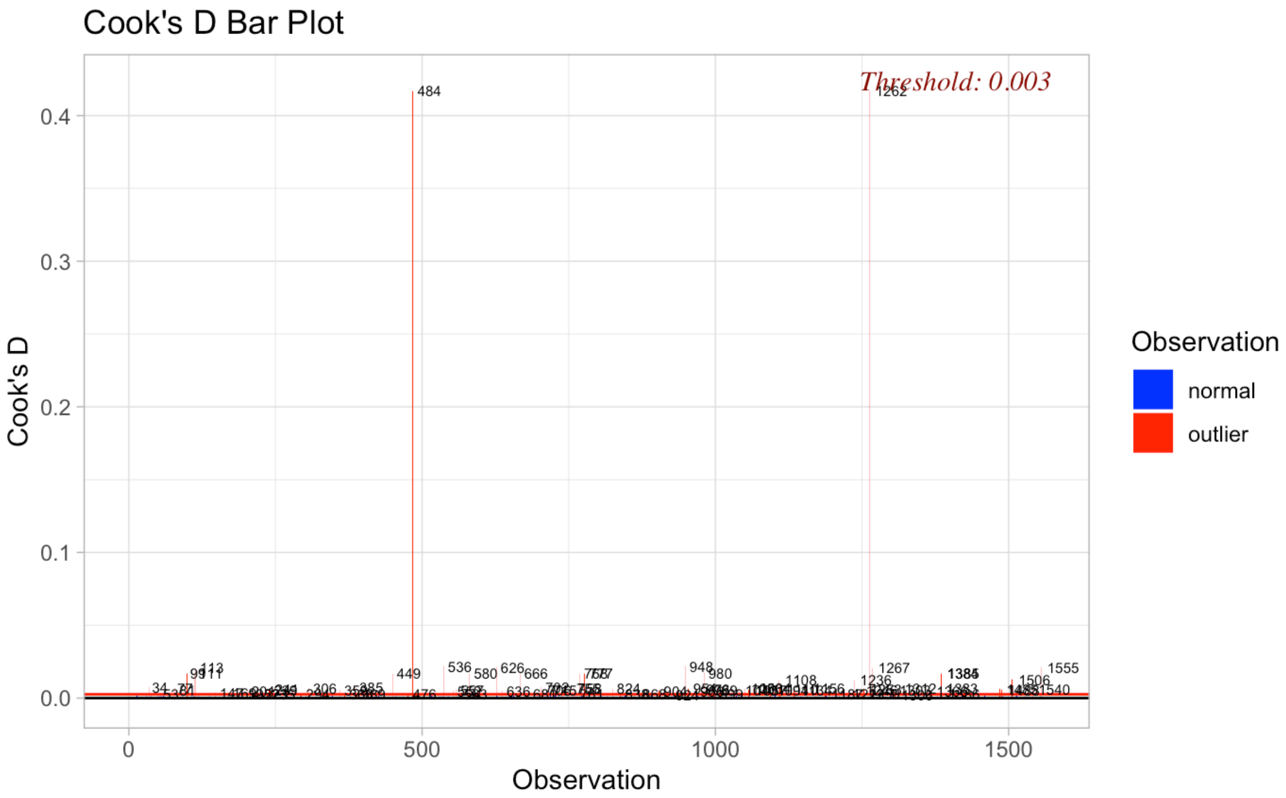
Base Model Cook’s Distance



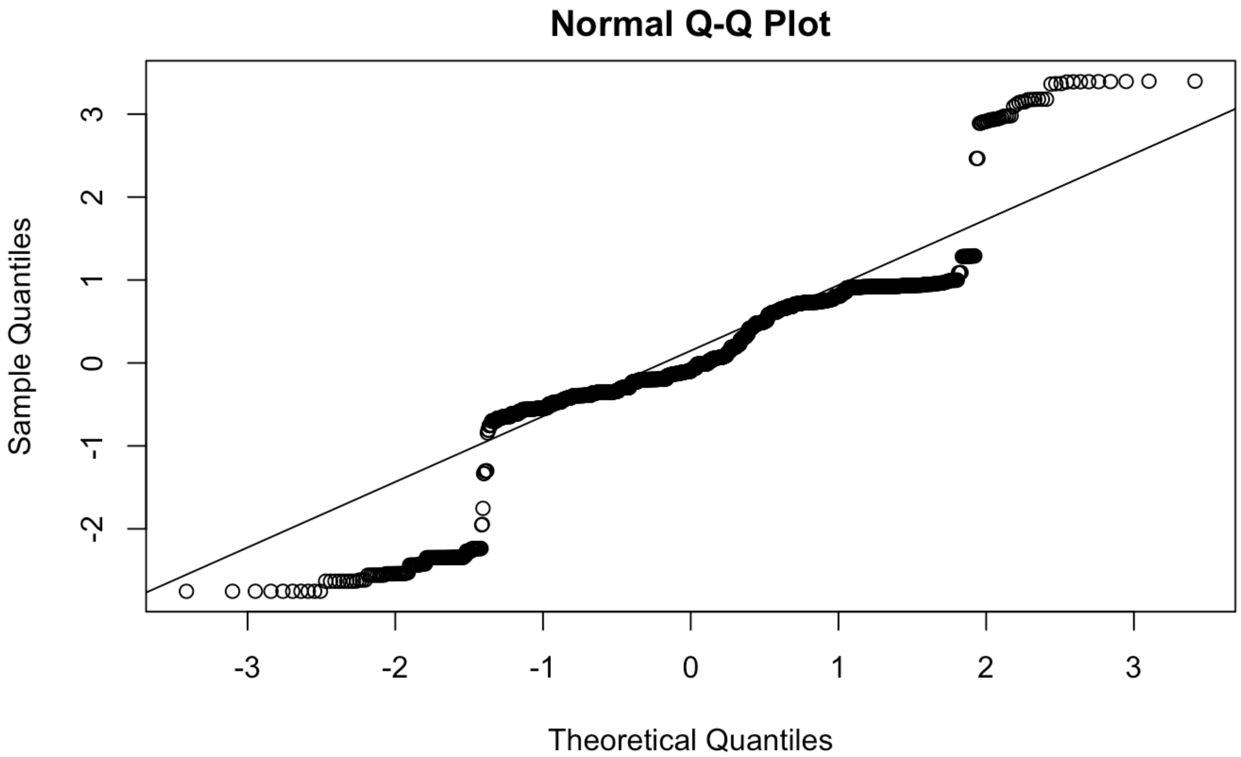
Base Model QQ Plot



Transformed Model Cook’s Distance



Transformed Model QQ Plot



Conclusion and Future Works

Overview

- Created multiple variations of linear model to predict garbage on beach
- With limited explanatory power, focused efforts on inferential data analysis and increasing model accuracy
- Created multiple visualizations during EDA to show relationships, or lack of relationships, between different variables
- Did not create model for estimating labor required
    - Without large explanatory power, a labor estimation would prove to be useless and misleading

Challenges Deep-Dive

Challenge 1

Lack of Explanatory Data

Model evaluations statistics show model does not explain a lot of variance in the data

- Tried to increase model accuracy but still need more data

Challenge 2

Sparsity of U.S. Data

With only 2,000 rows of data from specific regions, saw a lot of outliers

- Without other data sources available, tried to explain entries we did have

Challenge 3

Missing Features for Sub-Goals

Most of our sub-goals tried to show relationship with variables not in dataset

- Had to manually find more data and merge frames

Future Works

1. Create model with global data to reduce sparsity of data
    a. With 55,000 rows, may be able to create more accurate model

2. Obtain more external data and create new features
    a. Data on holidays or festivals celebrated in same county as beach may be a good predictor of more people going to beach
    b. Weather data
    c. Garbage cans on beach
    d. Type of plastics collected

3. Create estimation for number of volunteers needed to clean beach

Thank you!