Brandon Bennitt - A20416733
Yashwanth Praveen Pasupuleti - A20473431

CSP 571 Project Outline

Predicting the amount of workforce needed for coastline debris management

1. Project Proposal

- **A formal description of the project with a stated research goal**
  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. This sum can be drastically decreased 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. To take it one step further, given the amount of people needed to clean the area, organizations can estimate the amount of money they need to spend to clean each area.

  For this project, we wish to predict the amount of money needed to clean up a beach/coastline in a given city in the United States.

- **A specific question or set of questions that the project seeks to address**
  1. How much money is needed to clean up a beach/coastline in a given city?
  2. What state has the highest proportion of garbage to population ratio?
  3. Is there a relationship between the amount of garbage on a beach and the state that the beach is in?
  4. Given the amount of garbage on beaches in a given state, can we accurately predict the amount of garbage we would see on a beach in a neighboring state?
  5. In certain cities, is it more likely to find a higher percentage of plastic in the garbage collected?
  6. Is the amount of garbage on the shorelines related to the political party that was nominated in the previous presidential election?

- **A proposed methodology/approach to the analysis that will be performed**
  1. Import the data into RStudio to begin the process of data analysis.
  2. Clean the dataset to ensure the data can be used to accurately train the model. We are going to filter the data to look at only observations in the U.S. so we can minimize the problem at hand.

3. Perform exploratory data analysis and data visualization to better understand the relationship between the factors of interest. Explore and visualize different relationships to answer some of the sub questions this project wishes to answer.
4. Manually decide what factors look like they are good predictors.
5. Train the model with the features and response chosen, with the response being the amount of garbage on a given beach.
6. Create a second model that will estimate the amount of volunteers needed to clean up the garbage given the amount of garbage and size of the beach.
7. Based on the number of volunteers, we will calculate a price estimate for labor to clean a beach in a given city.

- **A metric or set of metrics which will measure analysis results**
    - When performing model selection, we will create a variety of different models and choose the best model based on the lowest root mean squared error (RMSE) since we are looking at a regression task.
    - We will improve the accuracy of the model analyzing F-stat and adjusted R^2.
    - Precision and recall will be used to determine the performance of the model.

2. Project Outline

Literature Review and Related Work:

- NAOO article:
    - Short summary: General article on what pollution is, related definitions, and some effects of marine pollution.
    - https://www.noaa.gov/education/resource-collections/ocean-coasts/ocean-pollution
- Oceans at MIT article:
    - Short summary: Article describing some of the research of Dr. Marcus Eriksen who has been a guest lecturer at top universities such as MIT. The paper the article summarizes talks about 5 years of observational work at sea where he collected data on pollution in the oceans. The article points out some interesting findings from Dr. Eriksen's research, and highlights how he thinks we may be able to solve the marine plastic problem.
    - http://oceans.mit.edu/news/featured-stories/269000-tons-plastic-ocean-now-dr-marcus-eriksen.html
- PLOS ONE:
    - The paper and abstract that was discussed in the Oceans at MIT article is found on this website. There are several figures that describe models that were made from the study as well as all research and results from 5 years of observational studies at sea exploring marine pollution. The main goal of the paper was to estimate how

much pollution is at sea and grouping the garbage into different categories as well.
    - https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0111913
- Plastic Pollution:
    - Summary: This is a website that took data from 4 different sources/papers and did some data visualization to answer a range of questions about plastic pollution in the environment. The data taken from other sources is all open access and graphs/plots can be used with citation. Each paper provides different findings given that the studies were focused around researching a slightly different area of marine pollution.
    - https://ourworldindata.org/plastic-pollution?utm_source=newsletter

- Plastic Pollution an Ocean Emergency:
    - Summary: Short paper on plastic pollution, its effects on marine wildlife, and an explanation for what can be done to reduce waste.
    - Link: https://www.researchgate.net/profile/Wallace-Nichols/publication/268187066_Editorial_Plastic_Pollution_An_Ocean_Emergency/links/54c622550cf256ed5a9c8f3c/Editorial-Plastic-Pollution-An-Ocean-Emergency.pdf

- Plastic Pollution of the World's Seas:
    - Short article on plastic pollution in the world's oceans that outlines the problem, governmental decisions impacting the pollution, and a way to move forward.
    - Link: https://www.nature.com/articles/s41467-018-03104-3.pdf

Data Sources

- Earth Challenge Integrated Data: Plastic Pollution (MLW, MDMAP, ICC) 2015-2018
    - Sourced from three citizen science marine litter projects, this is an interoperable global plastics dataset from 2015-2018. This data set is sourced from the European Environmental Agency's Marine Litter Watch Program (MLW), The National Oceanic and Atmospheric Administration's Marine Debris Monitoring and Assessment Project's accumulation report (MDMAP) and Ocean Conservancy's International Coastal Clean-Up (ICC) dataset. The data has been cleaned and analyzed with a common data schema. The data contains approximately 55,000 observations with features such as location of pollution, type of plastic, and total count of plastics along with many more.
    - Link with dataset and summary: https://globalearthchallenge.earthday.org/datasets/EC2020::data-earth-challenge-integrated-data-plastic-pollution-mlw-mdmap-icc-2015-2018/about

- Features - 16:
  - (X,Y) - Specifies the location coordinates.
  - SubCountry_L1_FromSource (State)
  - SubCountry_L2_FromSource (City)
  - TotalWidth_m - width of the beach (in meters)
  - TotalLength_m - length of the beach (in meters)
  - ShorelineName - name of the beach/general location
  - TotalVolunteers - Number of volunteers that helped with the recorded task
  - Year - year which the cleanup took place
  - MonthNum - number between 1-12 to represent the month the cleanup took place
  - Day - number between 1-31 to record 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/large island
  - LAND_RANK - The land type encoded in numerical values

Data processing and pipeline - cleaning, imputing, transformation, outlier detection, etc.

- Download dataset as CSV from original dataset website (4.7 MB)
- Standardize variable names in order to train the model easier
- Remove rows that have NA values associated with features of interest
- For input features that are numerical in the dataset but need to be used for cluster analysis, transform them into R factors
- Perform basic data exploration to understand the scale of the data and detect outliers

Data stylized facts - distributional analysis, clustering, dimensionality reduction, etc.

- We plan on referring to the reference source that has dozens of graphs and visual analyzations to brainstorm different ideas for how the data may be related.
- Clustering will be performed for the LAND_TYPE features so we can see if there is any relationship between the land type and the amount of garbage found per beach.
- Dimensionality reduction will be used if the model seems to be bloated if many of the possible features are chosen as significant

Model Selection

- For the main model, we are going to be using a regression model
- We will start by performing a simple linear regression and analyzing the results
- In order to attain better accuracy for the model, we will attempt to train more advanced models such as Ridge regression, Lasso regression, ensemble methods, etc.
- We will choose the best model based on the metrics that were mentioned above such as F-stat, Adjusted R^2, and test MSE

Software Tools

- R for all data analysis and model training
- The libraries that we need in order to perform the analyses would be “tidyverse (especially ggplot2 and dplyr), knitr, scales, tidytext, lubrid8, PASWR2”.
- Github for project management and source control.