Professional Development and Research Skills
CMM507 Coursework
Group 2: Plastic Pollution in Oceans

ALEXANDER RITCHIE, *1911218@rgu.ac.uk*;

GEORGIOS ORFANAKIS, 1903446@rgu.ac.uk;
KAREN JEWELL, 1415410@rgu.ac.uk;
ROSHI SHRESTHA, 1903445@rgu.ac.uk;
STUART WATT, 1501869@rgu.ac.uk

May 11, 2020

1 Problem Statement

1.1 Overview

Marine pollution is a major global issue which impacts the environment, economy and human health. Although marine pollution is caused by many different materials, plastics consist of 60-80% of the marine litter. [1] [2] [3] [4] [5] Plastics are synthetic organic polymers and their lightweight feature and durability make it very suitable to make a range of products we use in our everyday life.[6] [7] These same features are what makes plastic a major component of pollution, due to overuse and non-managed waste disposal systems internationally, contributing to 10% of the waste generated worldwide.[6] As the global production of plastic increases, so does the problem in the marine environment. Due to its buoyancy, plastic debris can be dispersed over long distances and accumulate on the shorelines, even polluting the most remote areas [8]. Jambeck et al.,[9] reported that in 2010 alone, between 4.8 million and 12.7 million metric tons of plastics entered the ocean. Plastics are everywhere in the marine environment and urgent action is required to mitigate this problem and reduce its harmful impact.[10] [11]

1.2 Motivation

The impact of plastic pollution on marine life has been reviewed extensively. [12] [13] [14] [15] Over 700 marine wildlife species are affected due to entanglement in plastic ropes or materials, and from ingestion of plastics in the ocean.[12] Over time, plastic disintegrates into microplastics and nanoplastics which are easily consumed by fish and enter the human food chain. Plastics have been found in a third of fish caught in the UK which included popular fish such as cod, haddock, and mackerel.[16] The impact and effects of plastic entering the human food chain are still being studied, but plastic toxicity and the occurrence of microplastics and nanoplastics in the water supply also directly impacts human health, in addition to the contamination of seafood.[11] [17]

Although plastic litter has been a major cause of marine pollution for a while, its seriousness has only been realised recently and reducing plastic pollution has become a global aim. Research on plastic pollution in marine environments has played a big role in efforts to reduce it as well as in raising awareness globally. In order to understand plastic pollution in marine environments and its effects in the long term, it is essential to keep collecting data on patterns of marine debris around the world. Effective monitoring of plastic debris is essential to reducing the abundance of plastic debris everywhere. In addition, monitoring the type, frequency and the source of the litter is also important for marine pollution prevention initiatives. Most monitoring is done by survey, where organisations and volunteers record the types and frequencies of litter observed on the shoreline.[18] To understand the depth of the problem, it is essential to understand the amount and composition of marine litter. This can help in applying various mitigation strategies.

1.3 Objectives

The main objectives of this project is outlined as follows:

- To review available literature on marine plastic problems and their impacts
- To look at a suitable dataset to understand the composition of marine plastic pollution
- To create a linear regression model to predict the relevant frequency of distinct plastic debris categories over time
- To present the derived results and conclusions

2 Research

An integral part of this report required a literature search to identify how researchers have been trying to monitor marine pollution and find the problems associated with it. Several studies have reported the abundance of plastic as marine litter through scientific survey and citizen science methods. A 12-year observation of coastal debris pollution using citizen science in Taiwan revealed that most debris items found were plastic. [19] 19 categories of debris items were collected during the clean-up events and the five most commonly recorded debris categories were: plastic shopping bags, plastic bottle caps, disposable tablewares, fishing equipment, and plastic drinking straws. In a study covering western Japan and the eastern coasts of Russia [20], it found that 55% to 93.4% of items over the Japanese shores were plastic, and the second most abundant item was resin pellet which is also a form of plastic. On the eastern Russian coast, plastic items were also the most abundant, contributing to approximately 55% of all litter, mostly of plastic fragments. The composition of litter was similar in the two countries, although the proportion of plastics was much higher in Japan. [20] Further along the Asian upper east coast, hard plastic and styrofoam were the dominant plastic types found on Korean beaches. On average, hard plastic and styrofoam comprised 32% and 48.5% of the total debris count respectively.

In an older study over the Caribbean region, the most common types of debris found on the Caribbean coast of Panama were plastic and styrofoam, with the plastics being household or consumer related. Styrofoam packing materials were also abundant and may have come from trans-shipment activities of Colon's Free Zone, household waste, or offshore activity [21]. A 2016/2017 annual study of 8 beaches in Tenerife in the Canary Islands also found that plastic was the most abundant litter there. They also reported that there was more accumulated plastic debris in remote beaches compared to the beaches near the city indicating the movement of debris. However, more long term study is required to understand the changes in the results reported over time. [22]

It was observed by the authors of this report, that there were variations in how studies of litter accumulation had been conducted. The variations are present in the time span of the research, the parts of the coast from which litter was collected, as well as in the categorisation of litter, which creates difficulty when researchers want to compare different studies. The plastic pollution problem essentially requires the ability to assess changes in accumulation rates and composition, trends over time, and the effectiveness of management systems, which is a hard task without good monitoring methodologies. Although monitoring of marine litter is currently carried out within a number of countries around the world, the methods of survey and monitoring used tend to be very different, preventing comparisons and harmonisation of data across regions or time.

This is why the scientific community has been trying to create some common ground, which has led to initiatives joined by many countries worldwide. One of these initiatives, and probably the most important, is the International Clean Coast (ICC) program which is a new, long term approach to cleaner beaches using various activities to increase public awareness. [23] This initiative aims to develop a comprehensive litter characterisation scheme that uses both material composition and form. This allows Litter Monitoring Repeated (LMR) surveys of beaches, seabeds and/or surface waters to determine litter quantities such that information can be compared with baseline data to identify if changes occur over time or in response to management arrangements.

The ICC uses specifically developed categorisations of coast litter, with the most accepted being the Clean Coastal Index (CCI) protocol specific to the operational clean-up of beaches, which is useful for its simplicity and information provided, allowing comparisons between different times and places. The CCI is the recommended tool for evaluation of actual coast cleanliness, measuring plastic debris as an indicator of beach cleanliness, precluding bias by the assessor. The CCI also proved to be a useful tool for measuring progress and the success of activities in raising awareness among the general public. [24]

A study in Israel using the CCI protocol for categorisation found that plastic was the most ubiquitous beach litter item. An important contribution to this study was the ability to compare its findings to other Mediterranean beaches, showing that plastic was the dominant pollutant in the region, and non-plastic litter being highly specific to the region and cannot be treated universally. [25] In another study on litter pollution in a region of India, once again following the CCI protocol for the categorisation of litter, found that plastic was the main form of litter at approximately 45% of total litter. Plastic bags topped the index at 33%, followed by food wrappers, then plastic cups, and cigarette/cigar tips amounting to 5.5%. [26] The use of the common protocol in these two studies allowed researchers to compare their findings and create common plastic pollution models, even though the two coasts are continents apart.

Another study conducted in Cadiz, on the other side of the Mediterranean from Israel, found that plastic bottles and containers were the most frequent littered items, followed by plastic bags. This research also pointed out that surveys are heavily affected by clean-ups performed at beaches. [27] Even though this study reaches important conclusions on the correct ways to clean coasts, it cannot be easily compared, or its conclusions easily applied without any standardised protocol.

From this review, it is evident that there have been many studies conducted to monitor marine pollution in various different ways. One of the most cost effective and efficient ways is the use of citizen science, where the general public can record any observations of marine litter [28] and has beens been successfully used in many studies. [29] [30] This method is also gaining popularity in marine pollution monitoring [31] [32] [33] [34] and has also been assessed as a tool to increase awareness of the marine litter problem. [35]

As discussed before, observations not following a standard protocol with proper guidance could be incomparable and rendered ineffective. The Marine Debris Tracker (MDT) is a joint initiative between National Oceanic and Atmospheric Administration (NOAA) Marine Debris Program (MDP) and the Universities of Georgia, North Carolina and South Carolina [36]. The MDT allows anyone to record the marine debris observed, using an application on a mobile phone. The only report using data from this application is a web-based report by Tablada in 2018 [36], where data analysis was performed on 8 years of the data mainly focused in North America. The study also concluded that plastic was the main type of debris recorded, with cigarettes being the top identified littered item. The subsequent sections of this report will aim to contribute to this body of knowledge, using the worldwide coastal littering dataset from the MDT, with an interest in identifying if plastic is indeed the most abundant litter of the worldwide dataset, which would be in agreement with the various studies discussed in this review. Following which, this report will analyse the distribution of subclasses within the plastics found, and explore if there could be a way to computationally monitor and assess the coastal littering problem, using the established CCI categorisation of litter.

3 Methods

3.1 Data Description

The data used in this report was gathered using secondary data collection methods only. The authors did not collect or create any new data using primary methods. The data was downloaded from the Marine Debris Tracker website (www.marinedebris.engr.uga.edu) on 19 February 2020. The dataset is composed of 363,368 global observations from the start of the program in 2010, to the latest available date at the time of download.

As discussed in the section before, the MDT is a citizen science project where organisations or individuals can record observations of marine debris using a mobile phone application. [36] The user records the observation using a structured form, and chooses the category of debris from a list provided but there are also multiple available fields to populate, including non-mandatory fields and some allowing free text entry. An example of a non-mandatory field are "Lists" which are customisable groupings usually for organisations to group their own collection of records.The table below describes the structure of the data.

Field	Description	Type
ListName	custom groupings of records	Non-mandatory
ListID	ID for ListName	Automated
ItemName	the category of debris	Mandatory
ItemID	ID for ItemName	Automated
LogID	unique ID for the observation	Automated
Quantity	the number of pieces of debris observed	Mandatory
Error radius	radius around the site within the error for reasonable doubt	Mandatory
Latitude	coordinates of the location where the observation was made	Mandatory
Longitude	coordinates of the location where the observation was made	Mandatory
Altitude	coordinates of the location where the observation was made	Mandatory
Location	text description of the location where the observation was made	Mandatory
Description	open text description for the observation	Non-mandatory
MaterialDescription	the material of the debris	Mandatory
MaterialID	ID for MaterialDescription	Automated
Timestamp	the date and time of observation	Mandatory

3.2 Data Pre-processing

```r
# Setting the environment
library(tidyverse)
library(purrr)
library(magrittr)
library(treemap)
library(hexbin)
library(mapdata)
library(viridis)
library(lubridate)
library(imager)
library(xtable)
library(dplyr)

# Loading the data
data <- list.files(path = "data/debris/", full.names = TRUE) %>%
  lapply(FUN = read_csv, col_types = "ififidddddcfcif") %>%
  reduce(rbind)

# Checking the size of the data
cat("rows: ", nrow(data),"; columns: ", ncol(data))

## rows:  363368 ; columns:  15
```

From download, the data was cleaned to prepare it for analysis. First, the **Timestamp** datetime information was converted into a date type format and renamed as variable **Time**

```r
# Replacing the column for time as a date data type, renaming it "Time"
data$Time <- data$Timestamp %>%
  parse_datetime(format = "%Y%m%d%H%M%S")
data$Timestamp <- NULL
```

The dataset was then inspected for missing values which were observed in the **Location** and **Description** fields, but as these are non-mandatory and open text fields it is to be expected, and all other fields are complete.

```r
# Identifying missing values
data %>% select_if(function(x) any(is.na(x))) %>% colnames()

## [1] "Location"    "Description"
```

It was observed that observations were incomplete for years 2010, 2011 and 2020, and so it was decided to retain only complete years of information for fair and standard comparisons.

```r
# Count the number of months in each year
data %>%
  mutate(year = year(Time),
         months = month(Time)) %>%
  select(year, months) %>%
  unique() %>%
  group_by(year) %>%
  summarize(nmonths = n())

## # A tibble: 11 x 2
##     year nmonths
##    <dbl>   <int>
## 1  2010       1
## 2  2011      10
## 3  2012      12
## 4  2013      12
## 5  2014      12
## 6  2015      12
## 7  2016      12
## 8  2017      12
## 9  2018      12
## 10 2019      12
## 11 2020       2
```

```
# Filter for observations occuring between the years 2012-2019 inclusive
data <- data %>% filter(as.integer(year(Time)) %in% 2012:2019)
```

```
# Checking the size of the data
cat("rows: ", nrow(data),"; columns: ", ncol(data))

## rows:  349556 ; columns:  15
```

Assessing the dataset for unique items it was observed that there are 8 material categories containing 55 different item subcategories and with them 7982 unique descriptions across the 349,556 observations.

```
# Counting unique values in fields
data %>%
  summarise_all(~length(unique(.))) %>%
  pivot_longer(cols = everything(),
               names_to = "Field",
               values_to = "Unique Values") %>%
  arrange(desc(`Unique Values`)) %>%
  xtable(caption = "The count of unique values in fields",
         label = "tab:unique",
         caption.placement = "top",
         floating=TRUE,
         type="latex",
         table.placement = "H")
```

It is difficult to determine exactly how many global locations are observed in this dataset as it is the combination of **Longitude**, **Latitude**, **Altitude** and/or **Location** which determines the global position of the observation. Yet in Figure 1 it is evident that the dataset is global, yet with a heavy influence of observations recorded in North America. It is also possible to identify that while most observed locations are on the coastline, some make reference to the middle of oceans and some are inland, which would account for pollution also observed in rivers and freshwater marine environments like lakes and natural reservoirs.

	Field	Unique Values
1	LogID	349556
2	Time	237066
3	Latitude	144820
4	Altitude	133174
5	Longitude	132316
6	Error Radius	16930
7	Description	7982
8	Location	1352
9	Quantity	494
10	ItemID	55
11	ItemName	55
12	Material ID	8
13	Material Description	8
14	ListID	1
15	ListName	1

Table 1: The count of unique values in fields

```r
# Map of all observed locations
world <- map_data("world")
data %>%
  select(Latitude, Longitude) %>%
  ggplot() +
    geom_polygon(data = map_data("world"), aes(x = long, y = lat, group = group), fill = "grey", alpha = 0.5) +
    geom_hex(aes(x = Longitude, y = Latitude), bins = 50) +
    scale_fill_viridis(trans = "log", breaks = c(5, 50, 500, 5000, 50000)) +
    theme_void() +
    guides(fill=guide_legend(title="Observations")) +
    theme(plot.title = element_text(size=10),
      text = element_text(size=8),
      panel.grid.major = element_blank(),
      panel.grid.minor = element_blank(),
      legend.position = "bottom")
```

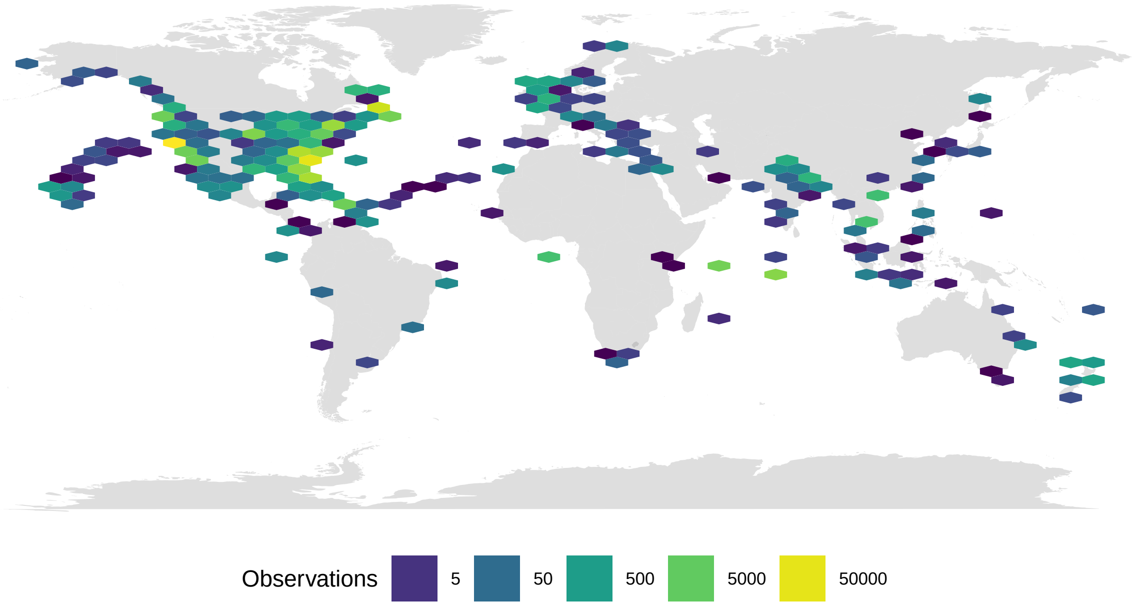


Figure 1: Observations by location

3.3 Quantities

In exploring the dataset, it was noted that there was a highly skewed distribution of **Quantities** recorded per observation.

```r
# Scatter plot showing distribution of observation quantities
data %>%
  mutate(Year = as.integer(year(Time))) %>%
  select(Quantity, Year) %>%
  group_by(Quantity, Year) %>%
  summarise(density = n()) %>%
  ggplot(aes(x=Quantity, y=density)) +
    geom_point() +
    scale_x_continuous(trans = 'log10', name='Quantity Recorded in Observation') +
    scale_y_continuous(name='Number of Observations') +
    theme_bw() +
    theme(text = element_text(size=8),
          panel.grid.major = element_blank(),
          panel.grid.minor = element_blank())
```

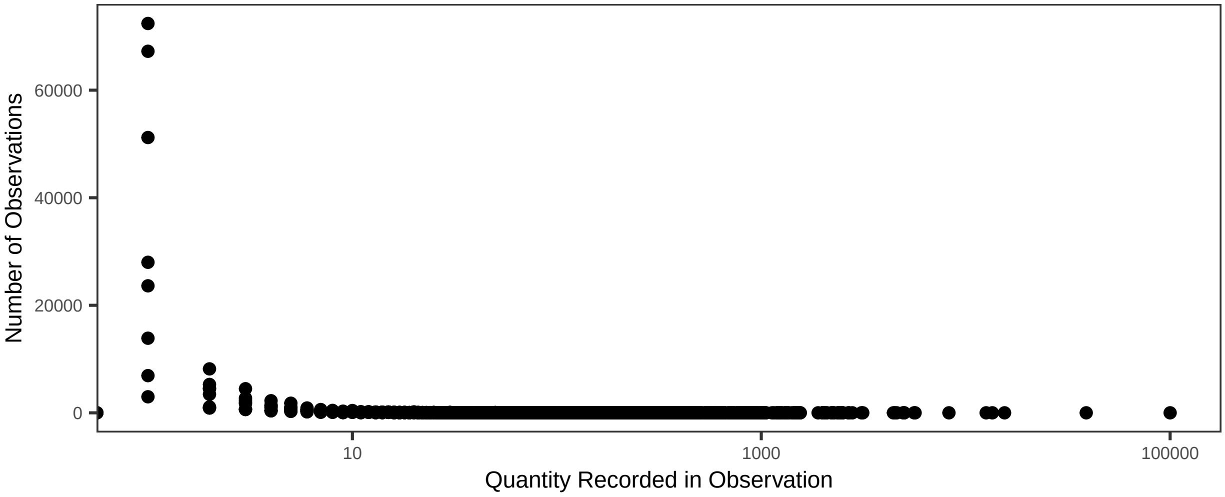


Figure 2: Distribution of quantities recorded by number of observations recording them

Figure 2 demonstrates the variance in the quantities by each observation in the dataset, and the frequency at which those quantities were recorded. It is easy to see the right-skew which indicates that the vast majority of the 300 thousand observations recorded quantities of 0 to 10, yet the minority record quantities varying between 10 and 100,000.

```r
# Identifying observations with the largest 10 quantities
data %>%
  mutate(Year = as.integer(year(Time))) %>%
  select(`Material Description`,ItemName,Year,Description,Quantity) %>%
  arrange(desc(Quantity)) %>%
  head(10)%>%
  xtable(caption = "Observations with the 10 largest quantities",
         label = "tab:top10quantities",
         caption.placement = "top",
         align = "lp{1.5cm}p{2.5cm}lp{7cm}l",
         floating=TRUE,
         type="latex",
         table.placement="H")
```

	Material Description	ItemName	Year	Description	Quantity
1	RUBBER	Tires	2019		100000.00
2	PLASTIC	Cigarettes	2012	Picked up during 20 minute beach cleanups over the course of 116 non consecutive cleanups	38875.00
3	PLASTIC	Cigarettes	2013		15496.00
4	PLASTIC	Plastic or Foam Fragments	2015		13500.00
5	PLASTIC	Cigarettes	2015		12594.00
6	PLASTIC	Cigarettes	2014		8281.00
7	PLASTIC	Cigarettes	2015		5656.00
8	PLASTIC	Cigarettes	2014		5598.00
9	PLASTIC	Cigarettes	2014	Earth and Surf Fest	5000.00
10	PLASTIC	Plastic or Foam Fragments	2019		5000.00

Table 2: Observations with the 10 largest quantities

```
# Identifying quantity percentiles
percent_quantity <-
  function(x){paste0(signif(length(data$Quantity[data$Quantity<=x])
                            /length(data$Quantity)*100, digits = 4), "%")}

tibble(Quantity = c("<=1","<=10","<=100"),
       ObservationProportion = c(percent_quantity(1),percent_quantity(10),percent_quantity(100))) %>%
  xtable(caption = "Proportion of observation having equal or less than the stated Quantity",
         label = "tab:pquantities",
         caption.placement = "top",
         floating=TRUE,
         type="latex",
         table.placement="H")
```

	Quantity	ObservationProportion
1	<=1	76.17%
2	<=10	96.18%
3	<=100	99.68%

Table 3: Proportion of observation having equal or less than the stated Quantity

Looking further into the distribution, 99.6% of all entries have observed quantities below 100, 96% as 10 or fewer and 76% as 1 item observed, suggesting that the vast majority of entries are citizen science entries. Interestingly some quantities are marked as 0, and there are a few observations with large quantities observed, possibly as a result of organised or research activity. In particular there are 5 observations with quantities of $> 10,000$ recorded in 2012, 2013, 2015 and 2019. It was decided not to exclude these high quantity observations as it would exclude a wealth of information gathered by organised groups or studies. Instead for the majority of this report, the count of observations is deemed to be a fairer indicator of density, rather than the quantity which is influenced heavily by the high quantity counts.

3.4 Categorisation

It is worth noting that of the 8 material categories, one is an ”Other” category which allows users to categorise an item as such, if it is not appropriate for any other option on the list.

```
# Listing of material categories
data %>% select(`Material Description`) %>%
  unique() %>%
  arrange(`Material Description`) %>%
  xtable(caption = "Listing of Material Descriptions",
         label = "tab:materials",
         caption.placement = "top",
         floating=TRUE,
         type="latex",
         table.placement="H")
```

	Material Description
1	PAPER & LUMBER
2	PLASTIC
3	METAL
4	CLOTH
5	OTHER ITEMS
6	FISHING GEAR
7	GLASS
8	RUBBER

Table 4: Listing of Material Descriptions

A data quality check performed on assessing the debris subcategories were unique by material type, highlighted that the subcategory ”Rubber Gloves” was associated with two material types, Plastic and Rubber. Yet otherwise, a one to many relationship exists between the parent material type and their item subcategories suggesting the integrity of the data remains intact.

```
# Checking for subcategory uniqueness
data %>%
  select(`Material Description`, ItemName) %>%
  distinct() %$%
  table(ItemName) %>%
  as_tibble() %>%
  filter(n > 1)

## # A tibble: 1 x 2
##   ItemName          n
##   <chr>         <int>
## 1 Rubber Gloves     2
```

Further investigation into the categorisation of Rubber Gloves revealed that the majority of Rubber Glove items were classified as plastic rather than rubber. When the observation descriptions were considered however, it revealed that the categorisations may be innaccurate, for example where an observation had ”Balloon” recorded in the description which is quite clearly not a glove. This raised the question of why these items were not classified as ”Other” materials but wrongly labelled as Rubber Gloves, which is perhaps a downside and reflective of the inconsistencies which can occur in citizen science data gathering.

```
# Checking material categorisation of Rubber Gloves
data %>% select(`Material Description`, ItemName, Quantity) %>%
  filter(ItemName == "Rubber Gloves") %>%
  group_by(`Material Description`) %>%
  summarise(Quantity = as.integer(sum(Quantity))) %>%
  xtable(caption = "Material categorisation of Rubber Gloves",
         label = "tab:2",
         caption.placement = "top",
         floating=TRUE,
         type="latex",
         table.placement="h")
```

	Material Description	Quantity
1	PLASTIC	2092
2	RUBBER	89

Table 5: Material categorisation of Rubber Gloves

```
# Checking the descriptions of Rubber Glove items
data %>% select(`Material Description`, ItemName, Description) %>%
  filter(ItemName == "Rubber Gloves", !is.na(Description)) %>%
  distinct() %>%
  xtable(caption = "Descriptions of Rubber Glove items",
         label = "tab:3",
         caption.placement = "top",
         floating=TRUE,
         type="latex",
         table.placement="h")
```

	Material Description	ItemName	Description
1	PLASTIC	Rubber Gloves	thermal
2	PLASTIC	Rubber Gloves	Near water
3	PLASTIC	Rubber Gloves	Taste of Omaha Cleanup
4	PLASTIC	Rubber Gloves	2 diff kinds
5	PLASTIC	Rubber Gloves	undefined
6	PLASTIC	Rubber Gloves	Latex
7	PLASTIC	Rubber Gloves	Hose
8	PLASTIC	Rubber Gloves	Cap
9	PLASTIC	Rubber Gloves	Vial
10	PLASTIC	Rubber Gloves	Gloves
11	PLASTIC	Rubber Gloves	Black intact
12	PLASTIC	Rubber Gloves	Clear intact
13	PLASTIC	Rubber Gloves	White glove fragmented
14	PLASTIC	Rubber Gloves	Fishing glove
15	PLASTIC	Rubber Gloves	Rubbery fragments
16	PLASTIC	Rubber Gloves	Rubber band and fishing bait
17	PLASTIC	Rubber Gloves	Balloon

Table 6: Descriptions of Rubber Glove items

```
# Line chart showing categorisation of Rubber Gloves over time
data %>%
  filter(ItemName == "Rubber Gloves") %>%
  mutate(months = floor_date(Time, 'month')) %>%
  group_by(months, ItemName,`Material Description`) %>%
  summarize(`Number of observations` = n()) %>%
  ggplot(aes(x = months, y = `Number of observations`)) +
    geom_line(aes(color=`Material Description`)) +
    theme_bw() +
    theme(panel.grid.major = element_blank(),
          panel.grid.minor = element_blank(),
          legend.position = "bottom")
```

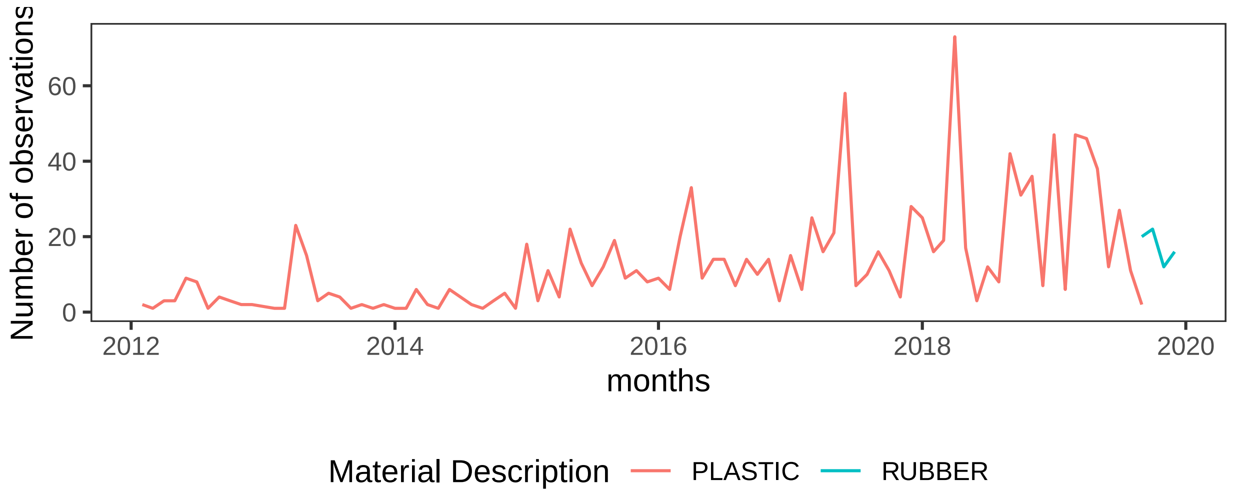


Figure 3: Categorisation of Rubber Glover over Time

However when the Rubber Glove categorisation was considered over time, it became apparent that in September 2019, the first use of ”Rubber” to classify Rubber Gloves appeared as did the the last categorisation as ”Plastic” at the same time. This suggests that the material classification was potentially changed at this time, which explains the two material categories and is not in fact an indicator of poor quality data.

3.5 Recategorisation

The **ItemName** subcategories of the Plastic type were recategorised following the CCI protocol to allow for consistency and comparability across studies as has been discussed in Section 2.

show original categories?

Figure 4: Debris by categorisation

```
# Recategorisation of all plastic items by CCI categories

# all cigarette related waste: 1, 4, 6, 22
# Food related waste: 3, 2,7,9,10, 17, 23, 11
# Non food related waste: 8, 14, 15, 16, 18, 19, 21, 20
# Plastic bags and Styrofoam packaging:12, 13
# Fragments: 5, 23, 24,25

plastic %>%
  select(ItemID, ItemName, category)

## Error in eval(lhs, parent, parent):  object 'plastic' not found

recategorise <- function(x){
  out = ""
  if(x %in% c(3,15,9,22)){out = "Cigarette related waste"}
  if(x %in% c(6,11,8,5,10,4,23,14)) out = "Food related waste"
  if(x %in% c(17,7,13,21,1,16,20,19)) out = "Other"
  if(x %in% c(2,18)) out = "Plastic bags and Styrofoam packaging"
  if(x %in% c(12,23,24,25)) out = "Fragments"
  if(out == "") stop(paste("Error in recategorise:", x))
  return(out)
}

plastic_types <- data %>%
  filter(`Material Description` == "PLASTIC") %>%
  select(ItemName, ItemID) %>%
  distinct() %>%
  mutate(label = 1:n()) %>%
  mutate(category = purrr::map(label, recategorise)) %>%
  mutate(category = as_factor(as.character(category))) %>%
  select(ItemID, category)


plastic <- data %>%
  filter(`Material Description` == "PLASTIC") %>%
  full_join(plastic_types, by = "ItemID")
```

```
plastic %>%
  mutate(month = month(Time, label = FALSE),
         year = as.integer(year(Time))) %>%
  group_by(month, year, category) %>%
  summarise(`Total Quantity` = sum(Quantity)) %>%
  ggplot(aes(x = month, y = `Total Quantity`, fill = category)) +
    geom_col(colour = "black", size = 0.2, position = "fill") +
    facet_wrap(~year, nrow = 2) +
    scale_fill_viridis(discrete = TRUE, option = "plasma") +
    xlab("Month") +
    ylab("Proportion of Items") +
    scale_x_continuous(breaks = 1:12) +
    theme(panel.grid.major.x = element_blank(),
          panel.grid.minor.x = element_blank(),
          legend.position = "bottom",
          legend.text=element_text(size=5)) +
    guides(fill=guide_legend(title="Category"))
```

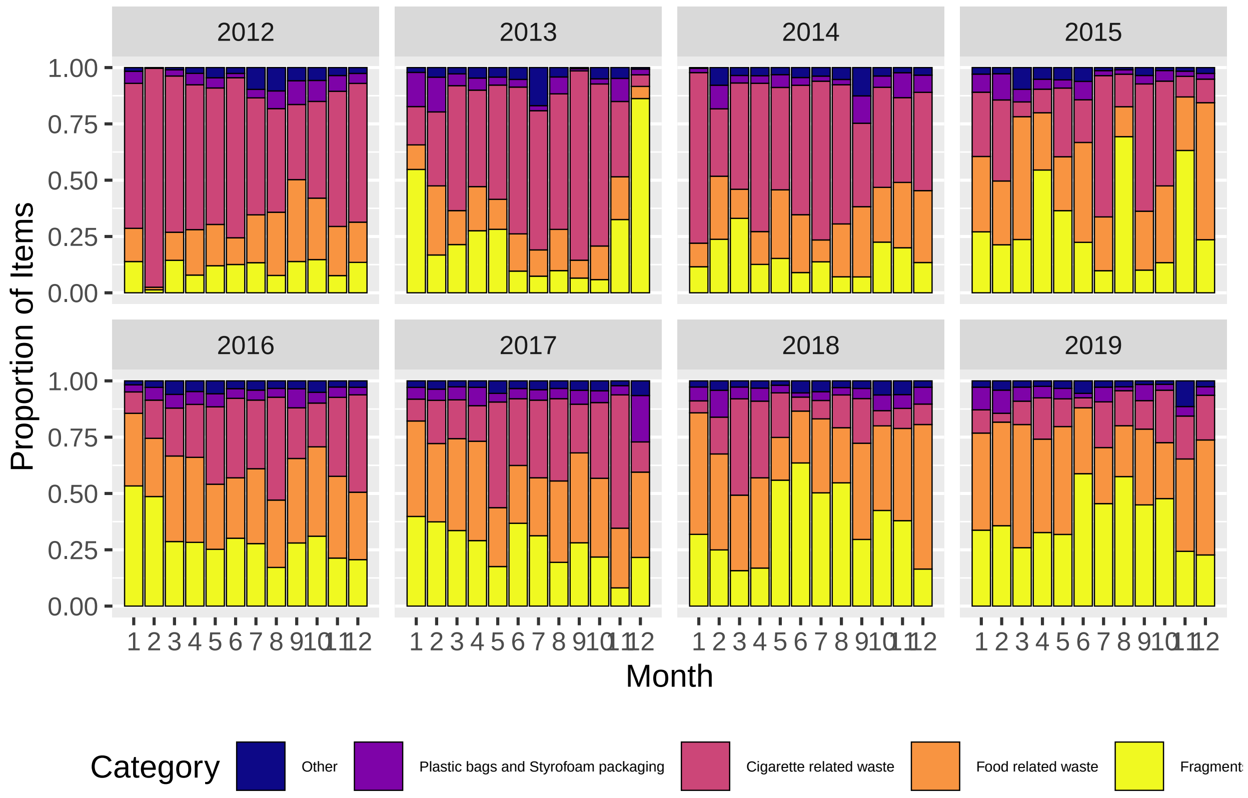


Figure 5: Relative frequencies of observed plastic waste by category

4 Experiments

4.1 Proportion Trends

After cleaning and recategorisation of the data, it was analysed to determine how pollutant proportions change over time.

```r
# Line chart of observations: Total v Plastic
data %>%
  mutate(Type = if_else(`Material Description` == "PLASTIC", "Plastic", "Other"),
         months = floor_date(Time, 'month')) %>%
  group_by(months, Type) %>%
  summarize(`Number of observations` = n()) %>%
  ggplot(aes(x = months, y = `Number of observations`, colour = Type)) +
    geom_line() +
    theme(legend.position = "bottom")
```

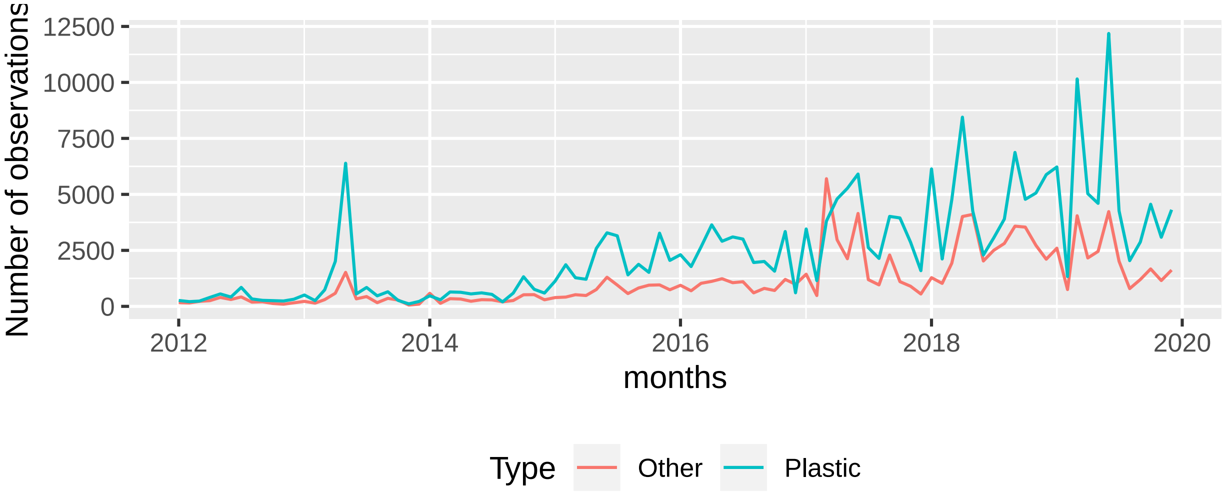


Figure 6: Observations of plastic debris v all debris

In this chart we see peaks and troughs in the number of observations - this is indicative of an increase in user activity. Seeing the peak in 2013 - that's not to say pollution suddenly increased in one period and disappeared, but monitoring activity spiked in that period. do we know why? We see more of these spikes in 2017, 2018 and 2019 but generally it is an increasing trend in monitoring activity.

Per the data, we see an increase in debris observed over time. Caution must be exercised in interpreting this information, as this increase may be reflective indeed of increased pollution levels, but it could also be because of increased activity in the citizen science project as it matures and gains a larger user base.

```r
# Column plot of average daily observations by year
data %>%
  mutate(Date = date(Time)) %>%
  full_join(tibble(Date = full_seq(c(min(date(data$Time)), max(date(data$Time))), period = 1)), by = "Date") %>% #
  group_by(Date) %>%
  summarise(Observations = length(LogID)) %>%
  ungroup() %>%
  mutate(Year = lubridate::year(Date)) %>%
  group_by(Year) %>%
  summarise(Average_Daily_Observations = mean(Observations)) %>%
  ungroup() %>%
  ggplot(aes(x = Year, y = Average_Daily_Observations)) +
    geom_col() +
    xlab("Year") +
    ylab("Average Daily Observations") +
    scale_x_continuous(breaks = 2012:2019) +
    theme_bw() +
    theme(text = element_text(size=8),
          panel.grid.major = element_blank(),
          panel.grid.minor = element_blank())
```

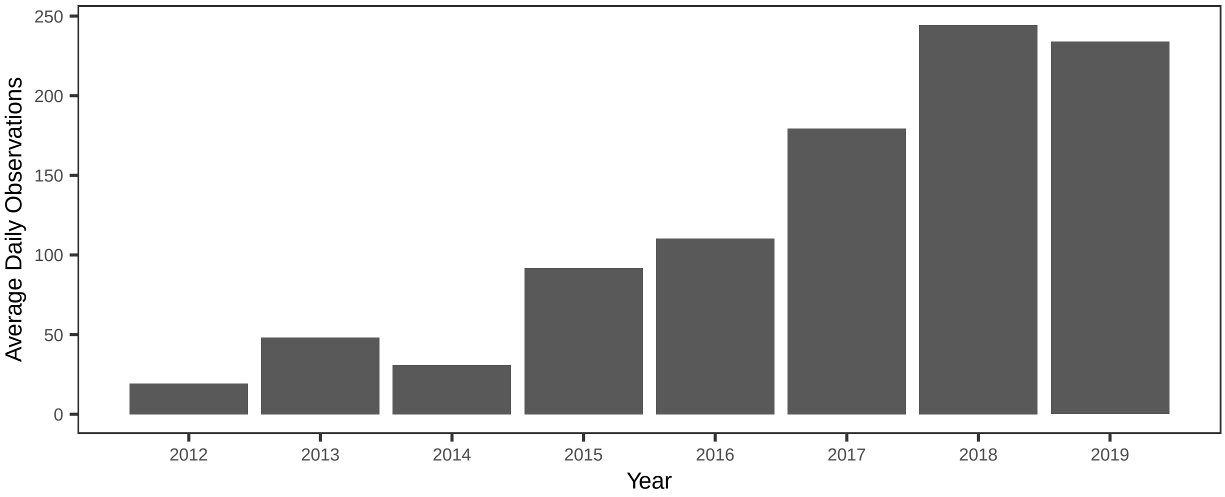


Figure 7: Average Daily Observations by Year

As we see in the chart above, an increase in the number of average observations per day lends support to the theory that the increase in observations is factored by increase in citizen science monitoring activity. That's not to say pollution has not also increased in this period, but it would be inaccurate to not account for the increase in monitoring activity.

can we put the above two charts together, side by side or top/bottom?

```r
# Column chart showing % distribution of material types
data %>%
  mutate(Year = lubridate::year(Time)) %>%
  group_by(Year, `Material Description`) %>%
  summarise(Counts = n()) %>%
  ungroup() %>%
  ggplot(aes(x = Year, y = Counts, fill = `Material Description`)) +
    geom_col(colour = "black", size = 0.2, position = "fill") +
    scale_fill_hue(l=50, c=150) +
    scale_x_continuous(breaks = 2012:2019) +
    scale_y_continuous(labels=scales::percent, name="") +
    theme_bw() +
    theme(text = element_text(size=8),
          panel.grid.major = element_blank(),
          panel.grid.minor = element_blank(),
          legend.position = "bottom")
```

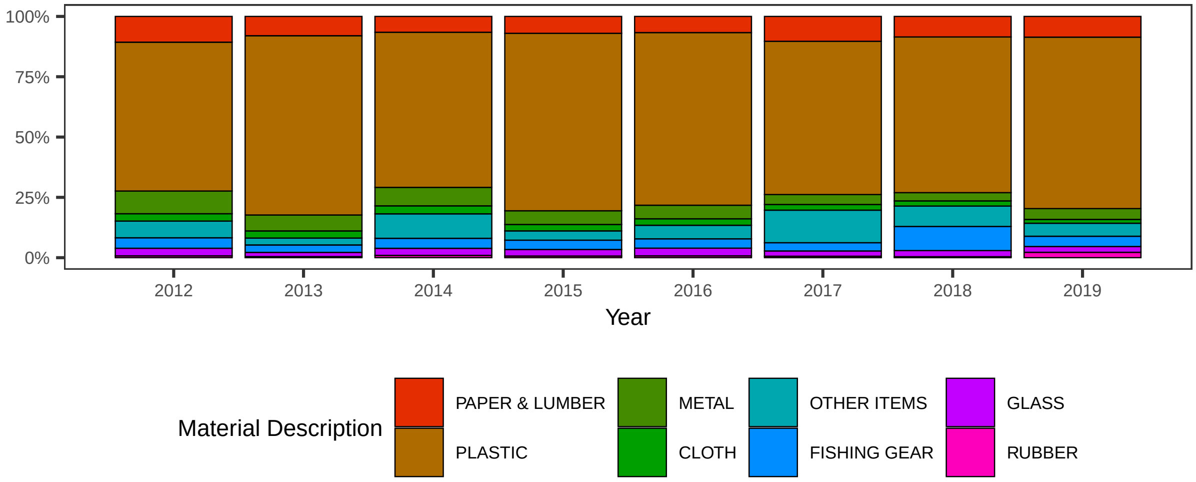


Figure 8: Proportion of Debris by Material Types

Analysis of the poportions of pollutant materials identifies that plastic is the dominant marine pollutant material, and its majority proportion of total debris is relatively consistent year on year. This could be due to an increase in new plastic pollution entering marine environments from source, not helped by the fact it is durable and does not decompose quickly, or that the other pollutant materials are decreasing. This analysis therefore is in agreement with the studies discussed in Section 2 that plastic is indeed the biggest problem in marine pollution accounting for approximately 70% of all marine pollution.

```r
# Line of just Plastic
data %>%
  mutate(Year = lubridate::year(Time)) %>%
  group_by(Year, `Material Description`) %>%
  summarise(total = n()) %>%
  mutate(prop = total / sum(total),lab=scales::percent(prop)) %>%
  filter(`Material Description`=="PLASTIC") %>%
  ggplot(aes(x = Year, y = prop, label = lab))+
  geom_text(nudge_y = 0.05, color = "black") +
  geom_line(aes(x = Year, y = prop))+
  geom_point(aes(x = Year, y = prop)) +
  scale_y_continuous(labels=scales::percent, name="", limits = c(0, 1)) +
  theme_bw() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        legend.position = "bottom")
```

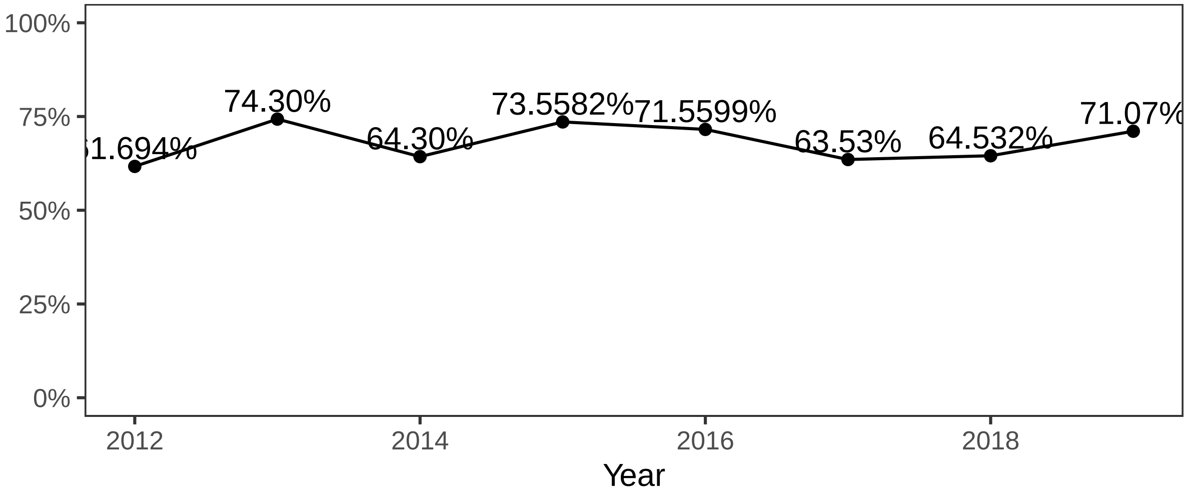


Figure 9: Trend of Plastic as a proportion of Total Debris

When considering the charts above, it can be determined that even as total debris observations increase, so do plastic observations which means plastics remain a consistently common feature of the observed derbis over time.

```
# Facet of categories
plastic %>%
  filter(`Material Description` == "PLASTIC") %>%
  mutate( year = as.integer(year(Time))) %>%
  group_by(year, category) %>%
  summarise(total = n()) %>%
  mutate(prop = total / sum(total),lab=scales::percent(prop)) %>%
  ggplot(aes(x = year, y = prop)) +
    geom_line() +
    facet_wrap(~category, nrow = 2) +
    theme(axis.text.y = element_blank(),
          axis.title.x = element_blank(),
          axis.title.y = element_blank())
```

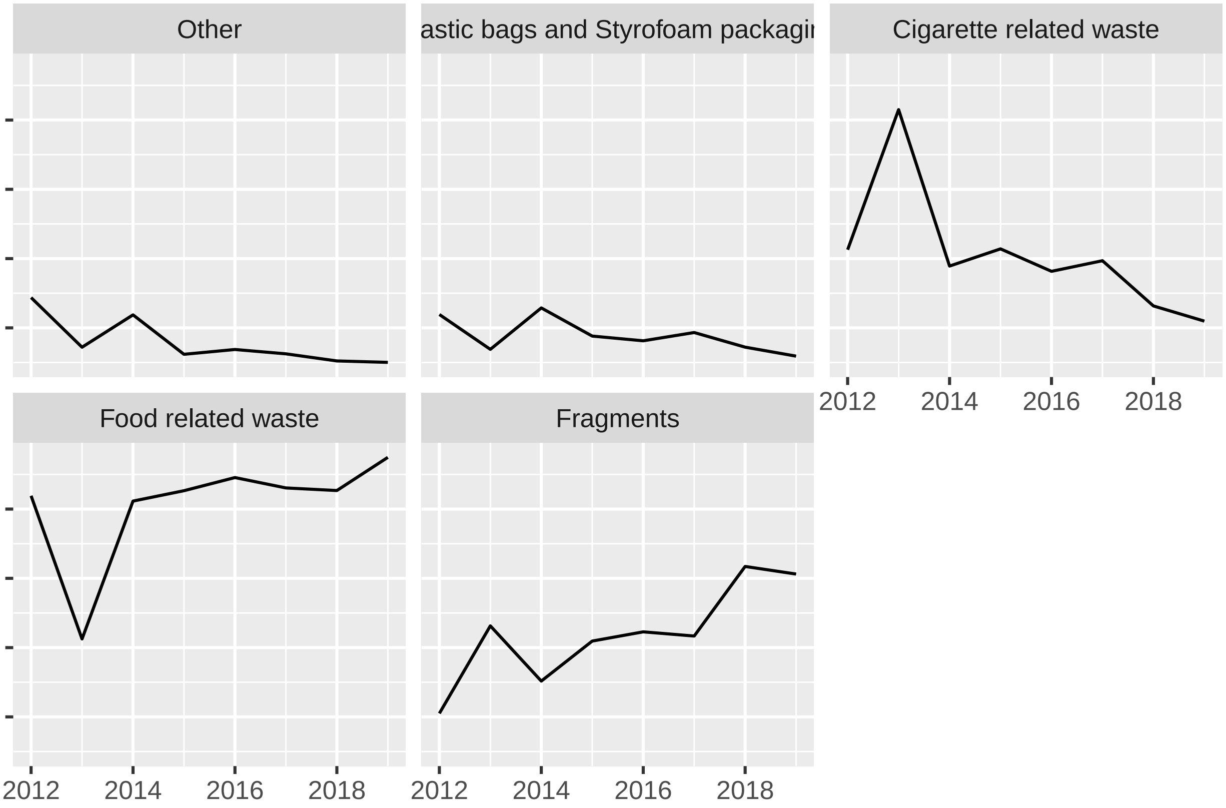


Figure 10: Proportion of Plastic Items Over Time

```
# Line chart of CCI categorised plastics
plastic %>%
  mutate(Year = lubridate::year(Time)) %>%
  group_by(Year, category) %>%
  summarise(total = n()) %>%
  mutate(prop = total / sum(total),lab=scales::percent(prop)) %>%
  ggplot(aes(x = Year, y = prop, color = category, label = lab))+
  #geom_text(color = "black") +
  geom_line(aes(x = Year, y = prop, color = category))+
  geom_point(aes(x = Year, y = prop, color = category)) +
  scale_y_continuous(labels=scales::percent, name="", limits = c(0, 1)) +
  theme_bw() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        legend.position = "bottom")
```

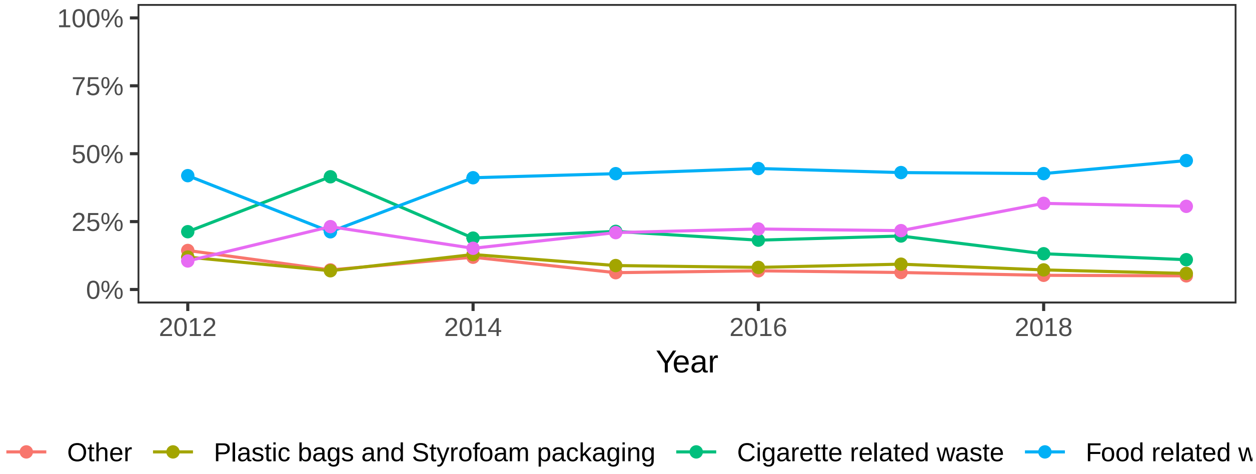


Figure 11: Trend of Plastic Proportions by CCI Categories

```
data %>%
  filter(`Material Description` == "PLASTIC") %>%
  mutate(year = as.integer(year(Time))) %>%
  group_by(year, ItemName) %>%
  summarise(total = n()) %>%
  mutate(prop = total / sum(total),lab=scales::percent(prop)) %>%
  ggplot(aes(x = year, y = prop)) +
    geom_line() +
    scale_y_continuous(labels=scales::percent, name="% of Total Plastic") +
    facet_wrap(~ItemName, nrow = 5) +
    theme(text = element_text(size=6),
          axis.text.x=element_text(angle=45, hjust=1),
          axis.ticks.y = element_blank(),
          axis.title.x = element_blank(),
          axis.title.y = element_blank())
```

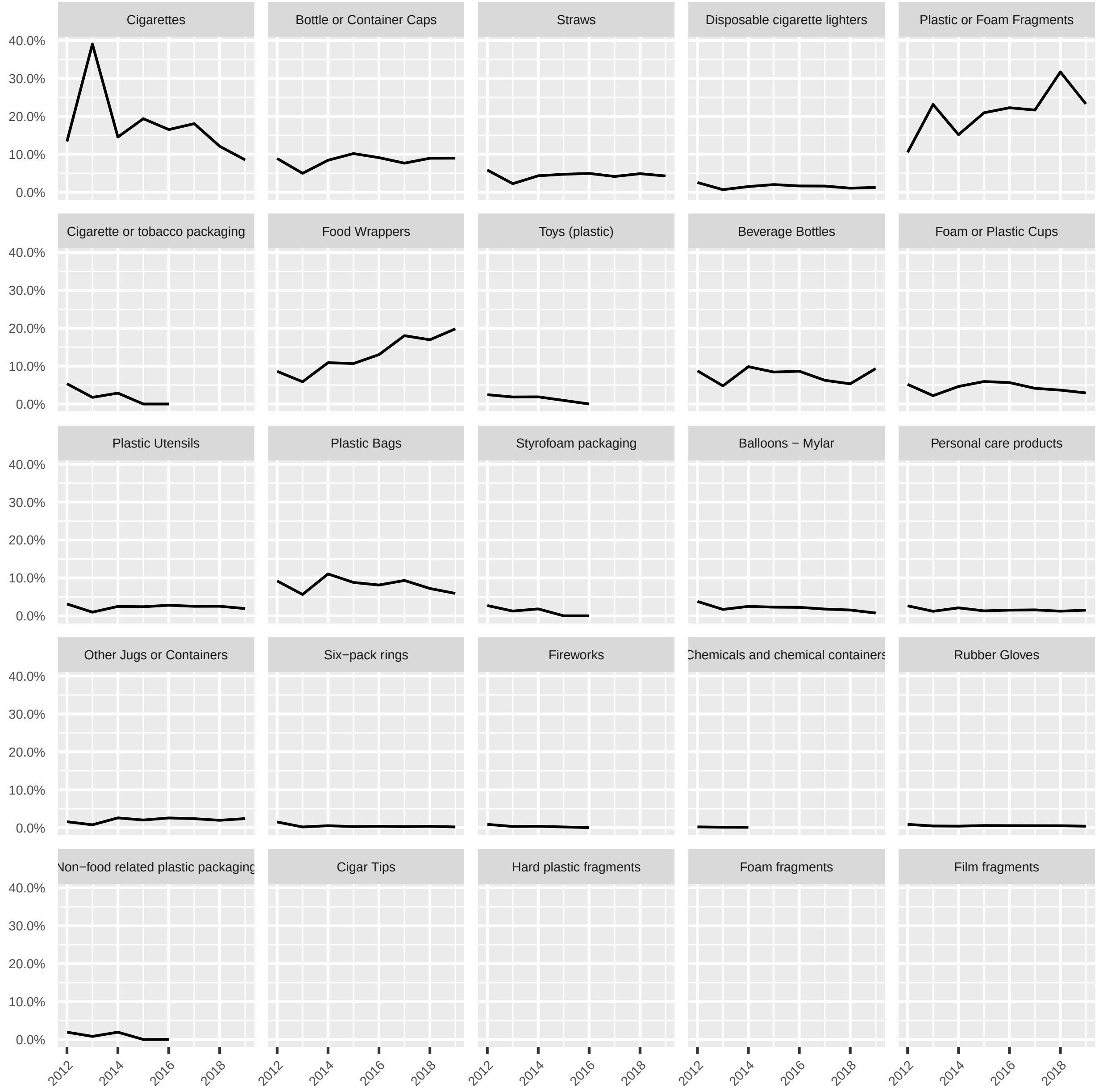
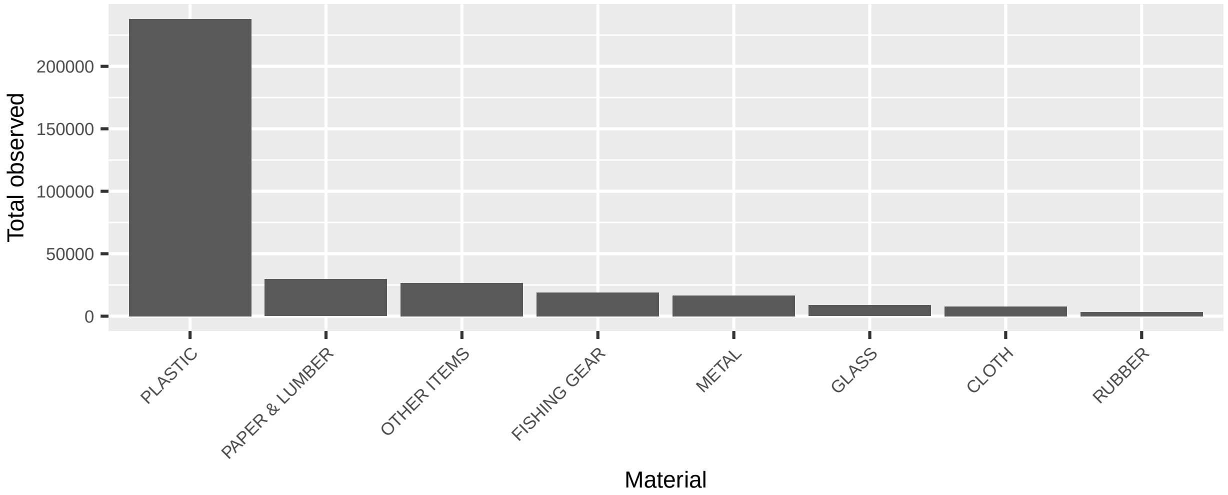


Figure 12: Proportion of Plastic Items Over Time

As these charts consider the proportion of the subcategory item of total plastics, it is important to note that the increase in Plastic or Foam fragments and Food Wrappers might be attributed actually to a decrease in other subcategories such as Cigarettes. Essentially, not that these items are becoming increasingly common, but that the share between them is levelling. Considering the discussions above, it should be recalled that all debris and plastic are being increasingly observed, so a decrease in a subcategory item is welcomed yet more must be done. To explain the decrease, it is possible that better waste management practices for cigarettes have been put into place, for example the introduction of penalties for literring cigarette butts or installing better cigarette butt collection points in high activity areas, resulting the decrease of observations witnessed here. It could however also be influenced by changes in social patterns such as persons smoking less or moving to reusable portable vaping devices which result is the lower consumption of single-use cigarettes and the resulting debris. Eitherway, being able to view key subcategories like this allows researchers to identify where to focus their efforts to ultimately achieve the aim of reducing pollution.

4.2 Distribution of observed debris:

```
# Column chart of Debris Quantity by Material Type
data %>% select(Description,`Material Description`) %>%
  group_by(`Material Description`) %>%
  summarise(Observed = n()) %>%
  ggplot(aes(x = reorder(`Material Description`, -Observed), y = Observed)) +
    geom_col() +
    ylab("Total observed") +
    xlab("Material") +
    theme(text = element_text(size=8),
        axis.text.x=element_text(angle=45, hjust=1),
        plot.title = element_text(size=10))
```



```
    #coord_flip()
```

Figure 13: Material Quantities

The most populated material class is Plastic. Note that this does not necessarily mean that plastic is the largest quantity of debris, just that the individual number of items categorised is largest.

```
# Treemap of debris categories
data %>%
  select(`Material Description`, ItemName, Quantity) %>%
  group_by(`Material Description`, ItemName) %>%
  summarise(Quantity = sum(Quantity)) %>%
  treemap(index = c("Material Description", "ItemName"),
          vSize = "Quantity", draw = TRUE) -> tm
```

Figure 14: Debris categorisation

```r
#bar of debris categories
data %>%
  select(`Material Description`, ItemName) %>%
  group_by(`Material Description`, ItemName) %>%
  summarise(Observed = n()) %>%
  arrange(desc(Observed)) %>%
  head(15) %>%
  ggplot(aes(x=reorder(`ItemName`, -Observed), y=Observed, fill=`Material Description`)) +
  geom_bar(stat="identity") +
  xlab("Debris") +
  ylab("") +
  #coord_flip() +
  theme(text = element_text(size=8),
        axis.text.x=element_text(angle=45, hjust=1),
        legend.text=element_text(size=5),
        legend.position = "bottom")
```

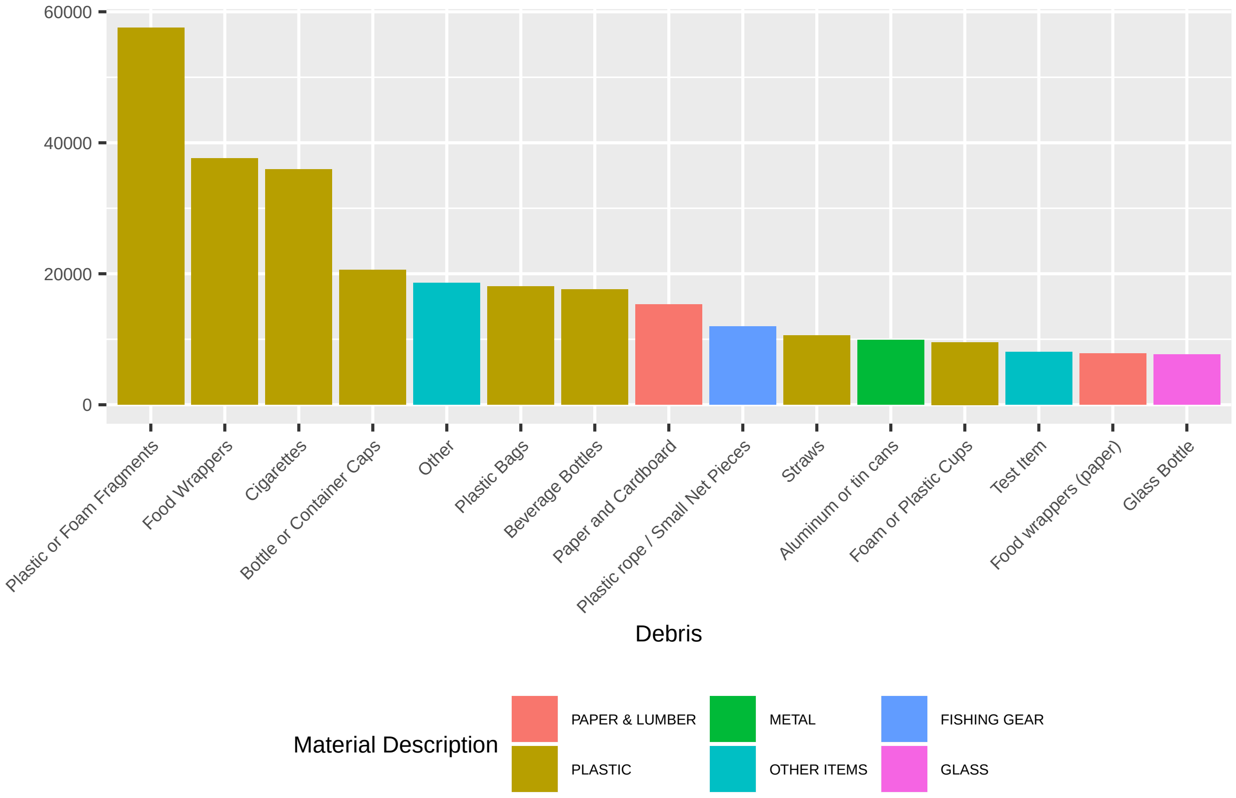


Figure 15: Top 15 Debris Items Observed between 2012 and 2019

Does this align with the other studies discussed in Section 2 where the main items identified were plastic bags, and in one study specifically from India which found cigarettes was actually their smallest pollutant at 5.5% [26].

4.3 Event-Driven Pollution

An interesting pattern which emerged when conducting exploratory analysis on the data, were peaks in the observations of debris classified as Fireworks consistently in July.

```
# Boxplot of fireworks distribution by month (across all years)
data %>%
  filter(`Material Description` == "PLASTIC",
         ItemName %in% c("Fireworks")) %>%
  mutate(month = month(Time, label = TRUE),
         year = as.integer(year(Time))) %>%
  group_by(month, year) %>%
  summarise(Observed = n()) %>%
  ggplot() +
    geom_boxplot(aes(x = month, y = Observed)) +
    xlab("Month") +
    ylab("Observed") +
    theme_bw() +
    theme(text = element_text(size=8),
      panel.grid.major = element_blank(),
      panel.grid.minor = element_blank(),
      legend.position = "bottom")
```

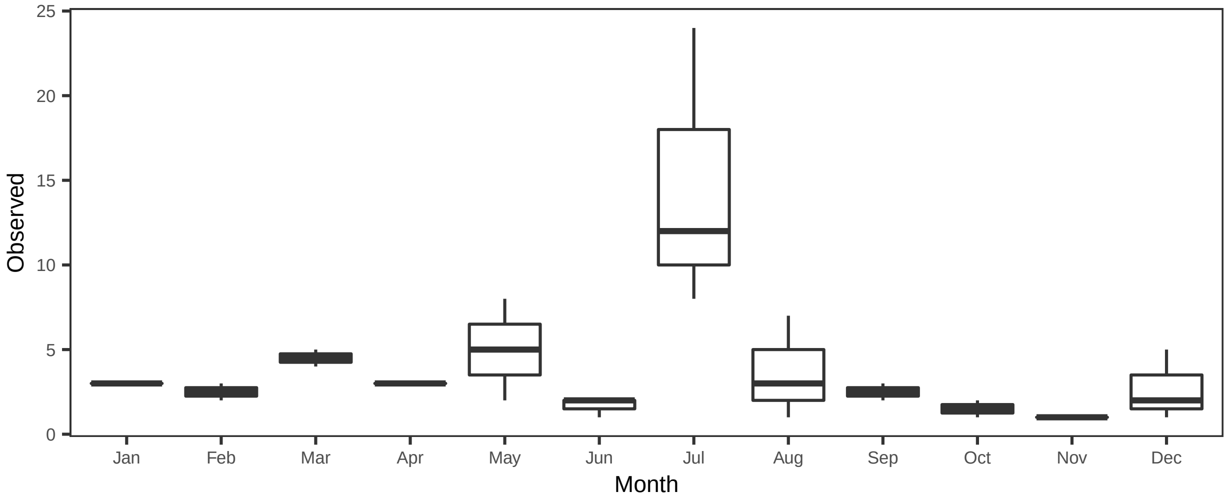


Figure 16: Firework debris 2012-2019

Given that it is known the dataset carries a heavy North American influence, a theory emerged that the peaks of observed debris were possibly related to American 4th of July celebrations. Looking at the locations of the firework observations in thevisualisation below certainly seemed to support that as firework related observations were made mostly in North America with a small sample in the United Kingdom. Interestingly, no spikes are observed in January which might suggest New years' celebrations, or in November for Bonfire Night celebrations on the 5th of November in the UK.

This gives rise to the idea that plastic pollution while sometimes blamed to be the effect of manufacturing processes, can also be largely driven by the disposable behaviours which humans encourage in these one off-events. Where the items are not intended for repeated use, used once and discarded, and eventually becoming pollutants in marine environments.

```r
# Map of locations having observed firework debris
data %>%
  filter(ItemName == "Fireworks") %>%
  select(Latitude, Longitude) %>%
  ggplot() +
    geom_polygon(data = map_data("world"), aes(x = long, y = lat, group = group), fill = "grey", alpha = 0.5) +
    geom_hex(aes(x = Longitude, y = Latitude)) +
    theme_void() +
    guides(fill=guide_legend(title="Observed")) +
    theme(plot.title = element_text(size=10),
      text = element_text(size=8),
      panel.grid.major = element_blank(),
      panel.grid.minor = element_blank(),
      legend.position = "bottom")
```

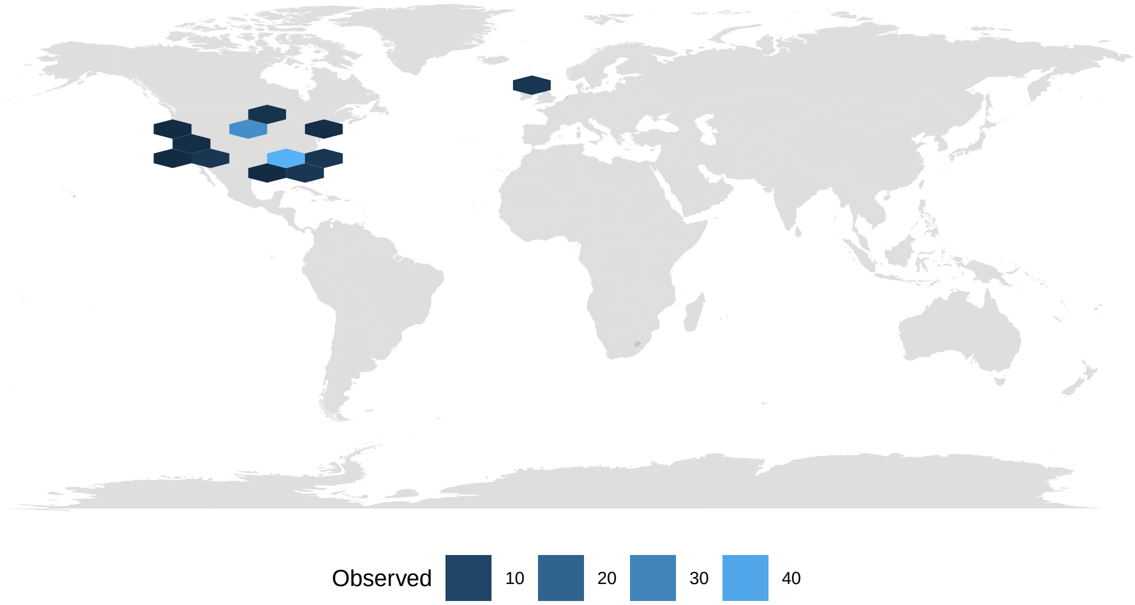


Figure 17: Boxplot of fireworks distribution by month, across all years

5 Predictive Modelling

Given the variability of plastic pollution trends given event-driven and location-driven pollution as explored earlier in this report, the authors of this report built a model to give more accurate predictions of expected pollution levels which can be used as a base model to assess the effectiveness of pollution reducing initiatives introduced by various organisations.

5.1 The Model

```
plasticN <- plastic %>%
  mutate(month = month(Time, label = FALSE), year = as.integer(year(Time))) %>%  group_by(year, category, month) %>
  summarise(`Total Quantity` = sum(Quantity))

####

df12N <- plasticN  %>%
  filter(year == 2012) %>%
  group_by(year, month) %>%
  mutate(freq = `Total Quantity` / sum(`Total Quantity`))

df13N <- plasticN  %>%
  filter(year == 2013) %>%
  group_by(year, month) %>%
  mutate(freq = `Total Quantity` / sum(`Total Quantity`))

df14N <- plasticN  %>%
  filter(year == 2014) %>%
  group_by(year, month) %>%
  mutate(freq = `Total Quantity` / sum(`Total Quantity`))

df15N <- plasticN  %>%
  filter(year == 2015) %>%
  group_by(year, month) %>%
  mutate(freq = `Total Quantity` / sum(`Total Quantity`))

df16N <- plasticN  %>%
  filter(year == 2016) %>%
  group_by(year, month) %>%
  mutate(freq = `Total Quantity` / sum(`Total Quantity`))

df17N <- plasticN  %>%
  filter(year == 2017) %>%
  group_by(year, month) %>%
  mutate(freq = `Total Quantity` / sum(`Total Quantity`))

df18N <- plasticN  %>%
  filter(year == 2018) %>%
  group_by(year, month) %>%
  mutate(freq = `Total Quantity` / sum(`Total Quantity`))

df19N <- plasticN  %>%
  filter(year == 2019) %>%
  group_by(year, month) %>%
  mutate(freq = `Total Quantity` / sum(`Total Quantity`))

dfTotN <- rbind(df12N, df13N, df14N, df15N, df16N, df17N, df18N, df19N)
```

```r
# plot for observing the data
(time_plotfr2N <- ggplot(dfTotN, aes(x = year, y = freq, color=category, fill = category)) +
  geom_smooth(method="lm", level=0.95) +
  theme_bw() +
  xlab("Years") +
  ylab("relative frequency") +
  expand_limits(y=0) +
  scale_y_continuous() +
  scale_x_continuous()+
  theme(legend.position="bottom")+
  theme(legend.text = element_text(size=5, face="bold")))
```

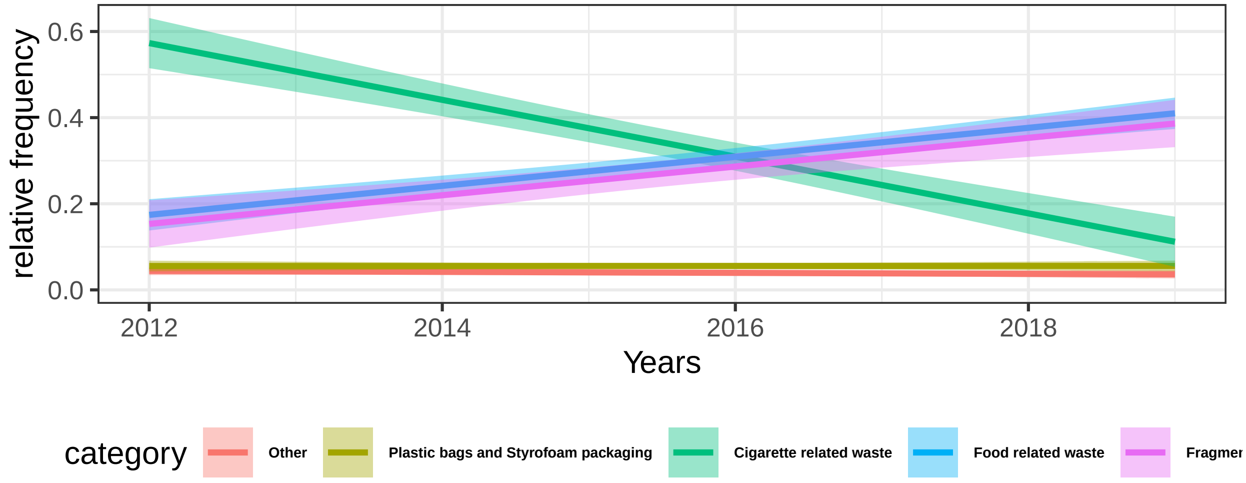


Figure 18: Portion of plastic

In this graphical representation of the relative frequency of the 5 different categories of plastic debris over the years, the insight to derive is that "Cigarette related waste, "Food related waste" and "Fragments" seem to experience some change, whereas "Other" and "Plastic bags and styrofoam packaging" seem to remain steady. A model is then created and tested on untrained data to see the early indications still stand.

```r
# create train and test set
n <- nrow(dfTotN)  # Number of observations
ntrain <- round(n*0.75)  # 75% for training set
set.seed(314)    # Set seed for reproducible results
tindex <- sample(n, ntrain)   # Create a random index
train_dfTotN <- dfTotN[tindex,]   # Create training set
test_dfTotN <- dfTotN[-tindex,]
```

```r
# modelling for category "Cigarette related waste"

train_Cigrel <- train_dfTotN %>%
  filter(category=="Cigarette related waste") %>%
  group_by(year)

test_Cigrel <- test_dfTotN %>%
  filter(category=="Cigarette related waste") %>%
  group_by(year)

set.seed(1234)
```

```
train_Cigrel.modelN <- lm(freq ~ year, data = train_Cigrel)
summary(train_Cigrel.modelN)

##
## Call:
## lm(formula = freq ~ year, data = train_Cigrel)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46278 -0.08200  0.01011  0.08529  0.39200
##
## Coefficients:
##               Estimate Std. Error t value         Pr(>|t|)
## (Intercept) 131.968708  15.355334   8.594 0.00000000000164 ***
## year         -0.065303   0.007618  -8.572 0.00000000000180 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1529 on 69 degrees of freedom
## Multiple R-squared:  0.5157,Adjusted R-squared:  0.5087
## F-statistic: 73.48 on 1 and 69 DF,  p-value: 0.0000000000018

print("PREDICTION")

## [1] "PREDICTION"

pred_Cigrel <- predict(train_Cigrel.modelN, test_Cigrel)
summary(pred_Cigrel)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1225  0.2531  0.3837  0.3706  0.4490  0.5796

actuals_predsCigrel <- data.frame(cbind(actuals=test_Cigrel$freq, predicteds=pred_Cigrel))
head(actuals_predsCigrel)

##     actuals predicteds
## 1 0.6937023  0.5796065
## 2 0.3334728  0.5796065
## 3 0.3284007  0.5143038
## 4 0.4283030  0.5143038
## 5 0.8405561  0.5143038
## 6 0.3343373  0.5143038

correlation_accuracy <- cor(actuals_predsCigrel)
min_max_accuracy <- mean(apply(actuals_predsCigrel, 1, min) / apply(actuals_predsCigrel, 1, max))
```

```
print(xtable(correlation_accuracy),table.placement="H")
```

	actuals	predicteds
actuals	1.00	0.64
predicteds	0.64	1.00

```
min_max_accuracy
```

[1] 0.6162189

```r
# modelling for category "Food related waste"

train_Foodrel <- train_dfTotN %>%
  filter(category=="Food related waste") %>%
  group_by(year)

test_Foodrel <- test_dfTotN %>%
  filter(category=="Food related waste") %>%
  group_by(year)

set.seed(1234)
train_Foodrel.modelN <- lm(freq ~ year, data = train_Foodrel)
summary(train_Foodrel.modelN)
```

```
##
## Call:
## lm(formula = freq ~ year, data = train_Foodrel)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.19587 -0.05990  0.00038  0.05769  0.33024
##
## Coefficients:
##              Estimate Std. Error t value     Pr(>|t|)
## (Intercept) -72.05047   11.04503  -6.523 0.0000000113 ***
## year          0.03590    0.00548   6.550 0.0000000102 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1002 on 66 degrees of freedom
## Multiple R-squared:  0.3939,Adjusted R-squared:  0.3848
## F-statistic:  42.9 on 1 and 66 DF,  p-value: 0.00000001015
```

```r
print("PREDICTION")
```

```
## [1] "PREDICTION"
```

```r
pred_Foodrel <- predict(train_Foodrel.modelN, test_Foodrel)
summary(pred_Foodrel)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1703  0.2421  0.2780  0.3062  0.3857  0.4216
```

```r
actuals_predsFoodrel <- data.frame(cbind(actuals=test_Foodrel$freq, predicteds=pred_Foodrel))
head(actuals_predsFoodrel)
```

```
##      actuals predicteds
## 1 0.18320210  0.1703163
## 2 0.28057222  0.1703163
## 3 0.27281324  0.1703163
## 4 0.16530055  0.2062113
## 5 0.07924182  0.2062113
## 6 0.14908896  0.2062113
```

```r
correlation_accuracy <- cor(actuals_predsFoodrel)
min_max_accuracy <- mean(apply(actuals_predsFoodrel, 1, min) / apply(actuals_predsFoodrel, 1, max))
```

```
print(xtable(correlation_accuracy),table.placement="H")
```

	actuals	predicteds
actuals	1.00	0.62
predicteds	0.62	1.00

```
print(min_max_accuracy)
```

[1] 0.7779188

```
# modelling for category "Fragments"
train_Frag <- train_dfTotN %>%
  filter(category=="Fragments") %>%
  group_by(year)

test_Frag <- test_dfTotN %>%
  filter(category=="Fragments") %>%
  group_by(year)

set.seed(1234)
train_Frag.modelN <- lm(freq ~ year, data = train_Frag)
summary(train_Frag.modelN)

##
## Call:
## lm(formula = freq ~ year, data = train_Frag)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23516 -0.09815 -0.02537  0.04237  0.66788
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -60.891268  16.153968  -3.769 0.000341 ***
## year          0.030346   0.008015   3.786 0.000323 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1486 on 69 degrees of freedom
## Multiple R-squared:  0.172,Adjusted R-squared:   0.16
## F-statistic: 14.33 on 1 and 69 DF,  p-value: 0.0003231

print("PREDICTION")

## [1] "PREDICTION"

pred_Frag <- predict(train_Frag.modelN, test_Frag)
summary(pred_Frag)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1643  0.2250  0.2857  0.2735  0.3464  0.3767

actuals_predsFrag <- data.frame(cbind(actuals=test_Frag$freq, predicteds=pred_Frag))
head(actuals_predsFrag)

##      actuals predicteds
## 1 0.01329086  0.1642836
## 2 0.12020997  0.1642836
## 3 0.13356847  0.1642836
## 4 0.07706507  0.1642836
## 5 0.07647059  0.1642836
## 6 0.54772432  0.1946293

correlation_accuracy <- cor(actuals_predsFrag)
min_max_accuracy <- mean(apply(actuals_predsFrag, 1, min) / apply(actuals_predsFrag, 1, max))
```

```
print(xtable(correlation_accuracy),table.placement="H")
```

	actuals	predicteds
actuals	1.00	0.57
predicteds	0.57	1.00

```
print(min_max_accuracy)
```

[1] 0.637178

```
# modelling for category "Other"
train_Other <- train_dfTotN %>%
  filter(category=="Other") %>%
  group_by(year)

test_Other <- test_dfTotN %>%
  filter(category=="Other") %>%
  group_by(year)

set.seed(1234)
train_Other.modelN <- lm(freq ~ year, data = train_Other)
summary(train_Other.modelN)

##
## Call:
## lm(formula = freq ~ year, data = train_Other)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.039992 -0.013652 -0.005996  0.009779  0.124826
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.442102   2.789048   0.876    0.384
## year        -0.001191   0.001384  -0.861    0.392
##
## Residual standard error: 0.02791 on 74 degrees of freedom
## Multiple R-squared:  0.009914,Adjusted R-squared:  -0.003466
## F-statistic: 0.741 on 1 and 74 DF,  p-value: 0.3921

print("PREDICTION")

## [1] "PREDICTION"

pred_Other <- predict(train_Other.modelN, test_Other)
summary(pred_Other)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.03717 0.03926 0.04253 0.04146 0.04313 0.04551

actuals_predsOther <- data.frame(cbind(actuals=test_Other$freq, predicteds=pred_Other))
head(actuals_predsOther)

##        actuals predicteds
## 1 0.0008478654 0.04551259
## 2 0.0279446331 0.04432144
## 3 0.0421825813 0.04432144
## 4 0.0500267953 0.04432144
## 5 0.0784077201 0.04313029
## 6 0.0350811369 0.04313029

correlation_accuracy <- cor(actuals_predsOther)  # 5.31%
min_max_accuracy <- mean(apply(actuals_predsOther, 1, min) / apply(actuals_predsOther, 1, max))
```

```
print(xtable(correlation_accuracy),table.placement="H")
```

	actuals	predicteds
actuals	1.00	0.14
predicteds	0.14	1.00

```
print(min_max_accuracy)
```

[1] 0.726022

```
# modelling for category "Plastic bags and Styrofoam packaging"
train_Plbag <- train_dfTotN %>%
  filter(category=="Plastic bags and Styrofoam packaging") %>%
  group_by(year)

test_Plbag <- test_dfTotN %>%
  filter(category=="Plastic bags and Styrofoam packaging") %>%
  group_by(year)

set.seed(1234)
train_Plbag.modelN <- lm(freq ~ year, data = train_Plbag)
summary(train_Plbag.modelN)

##
## Call:
## lm(formula = freq ~ year, data = train_Plbag)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.051045 -0.021221 -0.007823  0.007081  0.149541
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.0334621  3.6978320  -0.279    0.781
## year         0.0005404  0.0018346   0.295    0.769
##
## Residual standard error: 0.03539 on 72 degrees of freedom
## Multiple R-squared:  0.001204,Adjusted R-squared:  -0.01267
## F-statistic: 0.08677 on 1 and 72 DF,  p-value: 0.7692
```

```r
print("PREDICTION")

## [1] "PREDICTION"

pred_Plbag <- predict(train_Plbag.modelN, test_Plbag)
summary(pred_Plbag)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.05386 0.05440 0.05549 0.05563 0.05711 0.05765

actuals_predsPlbag <- data.frame(cbind(actuals=test_Plbag$freq, predicteds=pred_Plbag))
head(actuals_predsPlbag)

##      actuals predicteds
## 1 0.01901836 0.05386390
## 2 0.10543933 0.05386390
## 3 0.07005348 0.05386390
## 4 0.04430380 0.05386390
## 5 0.03557188 0.05440432
## 6 0.02274920 0.05440432

correlation_accuracy <- cor(actuals_predsPlbag)
min_max_accuracy <- mean(apply(actuals_predsPlbag, 1, min) / apply(actuals_predsPlbag, 1, max))
```

```r
print(xtable(correlation_accuracy),table.placement="H")
```

	actuals	predicteds
actuals	1.00	-0.14
predicteds	-0.14	1.00

```r
print(min_max_accuracy)
```

[1] 0.6891403

```r
plastic_category <-c("cigarette related waste", "food related waste","Fragments",
                      "Other","Plastic bags and Styrofoam packaging" )

slope_scores <- c(-0.063,0.038, 0.034, 0.000, 0.001)
slope_interpretation <-c("downward", "upward", "upward", "steady", "steady")
p_value<-c("<0.05","<0.05","<0.05", ">0.05",">0.05")
adjRsquared <- c(0.4416, 0.4324, 0.2177,  -0.0146, -0.01195)
corr_accuracy<-c(0.83, 0.50, 0.41, -0.36,-0.12)
min_max_Acc<-c(0.78,0.77,0.69, 0.64 ,0.61)

score_table1 <- data.frame(plastic_category, p_value,slope_scores, slope_interpretation, adjRsquared)
score_table2 <- data.frame(plastic_category, corr_accuracy, min_max_Acc)
```

```r
xtable(score_table1)
```

```r
xtable(score_table2)
```

	plastic_category	p_value	slope_scores	slope_interpretation	adjRsquared
1	cigarette related waste	<0.05	-0.06	downward	0.44
2	food related waste	<0.05	0.04	upward	0.43
3	Fragments	<0.05	0.03	upward	0.22
4	Other	>0.05	0.00	steady	-0.01
5	Plastic bags and Styrofoam packaging	>0.05	0.00	steady	-0.01

	plastic_category	corr_accuracy	min_max_Acc
1	cigarette related waste	0.83	0.78
2	food related waste	0.50	0.77
3	Fragments	0.41	0.69
4	Other	-0.36	0.64
5	Plastic bags and Styrofoam packaging	-0.12	0.61

5.2 Model Evaluation

On metrics presented on table:

$Pr(>|t|)$ is the p-value, defined as the probability of observing any value equal or larger than t if H0 is true. The larger the t statistic, the smaller the p-value. Generally, a 0.05 cutoff is used for significance. When p-values are smaller than 0.05, the hypothesis is rejected given that there is no significant difference between the means. If the p-value is larger than 0.05, it cannot conclude that a significant difference exists.

Correlation accuracy:
A simple correlation between the actuals and predicted values can be used as a form of accuracy measure. A higher correlation accuracy implies that the actuals and predicted values have similar directional movement, that is, when the actuals values increase the predicted values also increase and vice-versa.

MinMax Accuracy:
MinMax indicates how far off the model's prediction is. For a perfect model, this measure is 1.0. The lower the measure, the worse the performance of the model based on out-of-sample performance.in this case our model is....?

```
min_max_accuracy

## [1] 0.6891403
```

Adjusted R squared:
R-Squared gives the proportion of variation in the dependent (response) variable that has been explained by this model. Adjusted R-Squared is formulated such that it penalises the number of terms of the model.

On scores of metrics:

MinMax Accuracy is generally above 60%, but never exceeds 78% for all cases, which means that the model does a moderate job in predicting accurately the relative frequency of each category over time. The correlation accuracy is really good for ”cigarette related waste” category, but not that good for ”food related waste” and ”Fragments” categories, which implies that the predicted values do not always follow the true values observed in the same proportion. The categories ”Other” and ”Plastic bags and Styrofoam packaging”, where p-value is worse, score negative values which is troublesome. The higher the adjusted R squared metric the better. What is alarming here regarding this metric measure is the model created to predict for the last two categories, since it receives negative values.

In total we see that time is statistically significant in the change of proportions of certain categories of plastic waste: ”cigarette related waste”, ”food related waste”, ”fragments”, causing a downward, upward and upward movement respectively. This is not the case for the ”Other” and ”Plastic bags and Styrofoam packaging” categories, where though the models predicts a stagnation, there does not seem to be enough statistical evidence backing the credibility of the predictions of these two models.

It is important to notice that since it is a linear regression model, there exist no hyper-parameters for tuning, therefore no cross-validation comes in play. Since, only one predictor is used any kind of stepwise elimination is redundant. So, evaluation relies on metrics used.

5.3 Model Discussion

As seen in the study of debris in a Eastern Mediterranean coastal town [25] the variability of non-plastic litter composition should make us aware that the local context must be taken into account. Perhaps, given that non-plastic litter is a hugely varying category, this way of thinking could be expanded into the plastic debris category of ”Other”, given that it is inherently varied let alone in a dataset that gathers information over the entire world. In fact, this could in reverse explain the higher p-value, since high variation within the category studied would yield results of insufficient trustworthiness.

As mentioned in [37] many intertidal sites seem to be transit areas for debris, with exports matching imports over time. This could hugely impact a created model especially given that in the dataset we used we didn't have such information available. It is also important to acknowledge that our dataset is Northamerican centred which and given that it has information gathered over a period of 10 years we cannot be sure of categorization biase in the data, since there is no proof of a single sampling protocol followed. We have already spotted one such occasion, rubber gloves, and taken care of it, but further study on it is needed.

6 Conclusion and Future Work

As indicated by Jambeck et al.[9] 4-12 million tons of plastic waste generated worldwide will enter the marine ecosystem every year. Our results shows that there is no change in the percentage of marine debris that were recorded as plastic over time. However the recorded data are not in the same frequency every year and depend on the beach clean upsand national initiatives. Since these events aim mostly to clean-ups after big social events extra effort might have been made to retrieve entertainment based debris such as fireworks, food packaging and six pack rings. Our results confirm the fact that plastic is the most abundant litter in the marine environment globally. Plastic was included in the 68% of the observations recorded and it made up of over 70% of the quantity of debris that was recorded. Our results also indicate the successful use of citizen science projects to monitor marine pollution. There are various ways citizen science projects have been used for and here we focused on distribution and composition of marine litter on a global scale. Identifying the amount and type of debris can help in identifying the sources of pollution. This can then aid to find out ways to reduce the source in the first place. Event driven pollution such as fireworks also plays a role in marine pollution and this can be reduced by raising awareness amongst the public and trying to find more sustainable ways of celebrating events. The fact that plastic composition does not change over time is also quite alarming. Although there have been a lot of awareness programmes in the last decades, more needs to be done to see these changes that can reduce the damage done by plastic pollution.
Future work can include looking at the marine pollution debris in smaller scale by regions. Each region is different and there are various factors that can affect the data observations such as the frequency of data submissions are not at the same time everywhere. Since MDT is more focused on North America, this opportunity might be used to educate about it in other regions of the world. Although we will never be able to identify the real sources of pollution, it can help in raising awareness of the damage done by plastic pollution in the marine environment and make the public more environmentally conscious. Regulations to discourage single use plastic items and smoking in public places have been enforced in many countries. It would be interesting to look at the effect these are having on the marine debris pollution worldwide. However, more time and data will be required to see the change at a global scale.

Another line of study might be for example to compare the decreasing rate of cigarette use with the decreasing observation of cigarette debris which had been observed over the years to see if there is a correlation there which was hypothesised but never investigated due to not being within the direct inquiry of the report. It would also be useful to invest time and resources into a non-citizen science alternative to data collection in order to avoid the large spikes in observations during specifictimes such as world clean-ups day in order to get a more rounded dataset.

Lots of beach clean ups and litter collections are done in the beaches worldwide to estimate the marine pollution. However, the majority (70%) of the marine debris are not on the surface but on the seafloor. More research and studies are needed to accurately estimate the pollution in our oceans. Although marine pollution is a global issue, more needs to be done locally to reduce the sources of the pollution. In addition, standardised monitoring protocol and global partnerships are essential for efficient management of marine plastic pollution.

7 Project Management

7.1 Tools and Technologies

Group 2 communicated primarily using a dedicated Slack channel, with project materials managed on a Github repository, and weekly 1 hour in-person meetings. Slack was useful for making announcements; threaded conversations for topics if someone had a question; polling features which were used to collect votes, for example on deciding the topic of the research; but perhaps most importantly the integration with Github meant the group was automatically notified every time there was an update available for project materials, keeping members up to date at all times. Git controls on github meant document integrity was also maintained and there were no issues with persons working on materials which had become outdated and needed manual extraction and merging.

All project material used and the final report can be accessed from the *Public Github Repository*

In March 2020, the outbreak of the global Covid-19 pandemic and resulting lockdown measures enforced by the UK government, came into effect midway through this project. Fortunately, because online collaboration tools were already established and actively used, the "Stay at Home" order had little impact on the group's ability to complete work while remaining safely at home. Slack continued to be the main commmunication medium, with weekly 1 hour Skype calls replacing the original in-person meetings.

7.2 Project Progress

The group operated on a cyclical "divide and conquer" approch. In the weekly meetings, the scope for the week was agreed then divided into parts for group members to volunteer and adopt, working on it in the week. At the next meeting, the work was consolidated with individuals briefly presenting what they had achieved in the week, or what they may have found difficulty progressing. The newly converged scope was then again divided for the next week, so the cycle repeats ensuring each individual had a clear task for the week ahead and the whole group retained an awareness of what other work was being done and who was doing it. The discussions of each weekly meeting were also documented and hosted on the shared Github for all members to reference. This approach ensured there was continous progress throughout the project timeline while also allowing individuals the flexibility to manage their time around all other personal and professional commitments.

Table 7: Record of Team Meetings

No	Date	Topic	Alex	Georgios	Karen	Roshi	Stuart
1.00	2020-02-05	Group Formation: set up communication channel in Slack and GitHub repository	yes	yes	yes	yes	yes
2.00	2020-02-11	Agreed topic of ”Plastic Pollution”, distributed research activity for week	yes	yes	yes	yes	yes
3.00	2020-02-18	Presented inividuals' research findings and discussed hypothesis	yes	yes	yes	yes	yes
4.00	2020-02-25	Decided on final dataset to use and hypothesis of ”proportion of marine plastics pollution does not change over time”	yes	yes	yes	yes	yes
5.00	2020-03-04	Presentation draft agreed and agreed data needed re-categorising	yes	yes	yes	yes	yes
6.00	2020-03-10	Distributed section writing activity for week and discussed predictive model	yes	yes	yes	yes	yes
7.00	2020-03-17	Cancelled due to Covid-19 arrangements	yes	yes	yes	yes	yes
8.00	2020-03-24	Presentation dry run and literature sources distributed for review	yes	yes	yes	yes	yes
9.00	2020-03-31	Cancelled with agreement	yes	yes	yes	yes	yes
10.00	2020-04-07	Agreed structure of the final report	yes	yes	yes	yes	yes
11.00	2020-04-14	First review of the final report	yes	yes	yes	yes	yes
12.00	2020-04-22	Discussed consolidated changes from first review	yes	yes	yes	yes	yes
13.00	2020-04-28	Cancelled to support other modules	yes	yes	yes	yes	yes
14.00	2020-05-05	Cancelled to support other modules	yes	yes	yes	yes	yes
15.00	2020-05-08	Refresh of project material and set plan for remaining time till submission	yes	yes	yes	yes	yes
16.00	2020-05-09	Reviewed coursework guidelines to ensure compliance and agreed on internal deadline	yes	yes	yes	yes	no
17.00	2020-05-10	Finalised sections and identified last gaps to address	yes	yes	yes	yes	yes
18.00	2020-05-11	Final review and congratulations all around on a successfully completed report	yes	yes	yes	yes	yes

7.3 Peer Assessment

Table 8: Peer Assessment out of 100

Peer.Review	Alex	Georgios	Karen	Roshi	Stuart
Alex	100	100	100	100	100
Georgios	100	100	100	100	100
Karen	100	100	100	100	100
Roshi	100	100	100	100	100
Stuart	100	100	100	100	100

The Peer Assessment Table presents the peer and self assement scores for all five of Group 2's members.

The tasks of the project was shared justly across the members of the group. Alex Aided in the research of appropriate papers for the project and writing sections for the final report. Georgios sourced for datasets, reviewed literature, wrote for the final report and lead in creating and assessing the linear regression model.Karen acted as a facillitator and document controller for the project, responsible for editing the final report and ensuring it compiled successfully. Roshi acted as lead for finding relevant literatures for the study, writing sections and editing the final report. Stuart sourced the dataset and lead in the exploratory data analysis, producing visualisations for the final report.

References

[1] José G.B Derraik. “The pollution of the marine environment by plastic debris: a review”. In: *Marine Pollution Bulletin* 44.9 (2002), pp. 842–852. ISSN: 0025-326X. DOI: https://doi.org/10.1016/S0025-326X(02)00220-5. URL: http://www.sciencedirect.com/science/article/pii/S0025326X02002205.

[2] Stephen Smith and Ana Markic. “Estimates of Marine Debris Accumulation on Beaches Are Strongly Affected by the Temporal Scale of Sampling”. In: *PloS one* 8 (Dec. 2013), e83694. DOI: 10.1371/journal.pone.0083694.

[3] Fan-Jun Kuo and Hsiang-Wen Huang. “Strategy for mitigation of marine debris: Analysis of sources and composition of marine debris in northern Taiwan”. In: *Marine pollution bulletin* 83 (Apr. 2014). DOI: 10.1016/j.marpolbul.2014.04.019.

[4] Peng Zhou et al. “The abundance, composition and sources of marine debris in coastal seawaters or beaches around the northern South China Sea (China)”. In: *Marine pollution bulletin* 62 (Sept. 2011), pp. 1998–2007. DOI: 10.1016/j.marpolbul.2011.06.018.

[5] Yong Chang Jang et al. “Sources of plastic marine debris on beaches of Korea: More from the ocean than the land”. In: *Ocean Science Journal* 49 (June 2014), pp. 151–162. DOI: 10.1007/s12601-014-0015-8.

[6] David Barnes et al. “Accumulation and fragmentation of plastic debris in global environments”. In: *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 364 (Aug. 2009), pp. 1985–98. DOI: 10.1098/rstb.2008.0205.

[7] Alex Sivan. “New perspectives in plastic biodegradation”. In: *Current Opinion in Biotechnology* 22.3 (2011), pp. 422–426. ISSN: 0958-1669. DOI: https://doi.org/10.1016/j.copbio.2011.01.013. URL: http://www.sciencedirect.com/science/article/pii/S0958166911000292.

[8] Jennifer Lavers and Alexander Bond. “Exceptional and rapid accumulation of anthropogenic debris on one of the world’s most remote and pristine islands”. In: *Proceedings of the National Academy of Sciences* 114 (May 2017), p. 201619818. DOI: 10.1073/pnas.1619818114.

[9] Jenna R. Jambeck et al. “Plastic waste inputs from land into the ocean”. In: *Science* 347.6223 (2015), pp. 768–771. ISSN: 0036-8075. DOI: 10.1126/science.1260352. eprint: https://science.sciencemag.org/content/347/6223/768.full.pdf. URL: https://science.sciencemag.org/content/347/6223/768.

[10] Lorena M. Rios, Charles Moore, and Patrick R. Jones. “Persistent organic pollutants carried by synthetic polymers in the ocean environment”. In: *Marine Pollution Bulletin* 54.8 (2007), pp. 1230–1237. ISSN: 0025-326X. DOI: https://doi.org/10.1016/j.marpolbul.2007.03.022. URL: http://www.sciencedirect.com/science/article/pii/S0025326X07001324.

[11] Chelsea M. Rochman. “The Complex Mixture, Fate and Toxicity of Chemicals Associated with Plastic Debris in the Marine Environment”. In: *Marine Anthropogenic Litter*. Ed. by Melanie Bergmann, Lars Gutow, and Michael Klages. Cham: Springer International Publishing, 2015, pp. 117–140. ISBN: 978-3-319-16510-3. DOI: 10.1007/978-3-319-16510-3_5. URL: https://doi.org/10.1007/978-3-319-16510-3_5.

[12] S.C. Gall and R.C. Thompson. “The impact of debris on marine life”. In: *Marine Pollution Bulletin* 92.1 (2015), pp. 170–179. ISSN: 0025-326X. DOI: https://doi.org/10.1016/j.marpolbul.2014.12.041. URL: http://www.sciencedirect.com/science/article/pii/S0025326X14008571.

[13] Susanne Kühn, Elisa L. Bravo Rebolledo, and Jan A. van Franeker. “Deleterious Effects of Litter on Marine Life”. In: *Marine Anthropogenic Litter*. Ed. by Melanie Bergmann, Lars Gutow, and Michael Klages. Cham: Springer International Publishing, 2015, pp. 75–116. ISBN: 978-3-319-16510-3. DOI: 10.1007/978-3-319-16510-3_4. URL: https://doi.org/10.1007/978-3-319-16510-3_4.

[14] Peter G. Ryan. “A Brief History of Marine Litter Research”. In: *Marine Anthropogenic Litter*. Ed. by Melanie Bergmann, Lars Gutow, and Michael Klages. Cham: Springer International Publishing, 2015, pp. 1–25. ISBN: 978-3-319-16510-3. DOI: 10.1007/978-3-319-16510-3_1. URL: https://doi.org/10.1007/978-3-319-16510-3_1.

[15] A.T. Williams and Nelson Rangel-Buitrago. "Marine Litter: Solutions for a Major Environmental Problem". In: *Journal of Coastal Research* 35.3 (2019), pp. 648–663. DOI: 10.2112/JCOASTRES-D-18-00096.1. URL: https://doi.org/10.2112/JCOASTRES-D-18-00096.1.

[16] "Occurrence of microplastics in the gastrointestinal tract of pelagic and demersal fish from the English Channel". In: *Marine Pollution Bulletin* 67.1 (2013), pp. 94–99. ISSN: 0025-326X. DOI: https://doi.org/10.1016/j.marpolbul.2012.11.028.

[17] Ana Markic et al. "Plastic ingestion by marine fish in the wild". In: *Critical Reviews in Environmental Science and Technology* 50.7 (2020), pp. 657–697. DOI: 10.1080/10643389.2019.1631990. eprint: https://doi.org/10.1080/10643389.2019.1631990. URL: https://doi.org/10.1080/10643389.2019.1631990.

[18] J.M. Coe and D.B. Rogers. "Marine Debris: Sources, Impacts, and Solutions". In: Environmental Management Series. Springer, 1997. ISBN: 9780387947594. URL: https://books.google.co.uk/books?id=aSoRAAAAYAAJ.

[19] Bruno A. Walther, Alexander Kunz, and Chieh-Shen Hu. "Type and quantity of coastal debris pollution in Taiwan: A 12-year nationwide assessment using citizen science data". In: *Marine Pollution Bulletin* 135 (2018), pp. 862–872. ISSN: 0025-326X. DOI: https://doi.org/10.1016/j.marpolbul.2018.08.025. URL: http://www.sciencedirect.com/science/article/pii/S0025326X18305897.

[20] Takashi Kusui and Michio Noda. "International survey on the distribution of stranded and buried litter on beaches along the Sea of Japan". In: *Marine pollution bulletin* 47 (Feb. 2003), pp. 175–9. DOI: 10.1016/S0025-326X(02)00478-2.

[21] Stephen D. Garrity and Sally C. Levings. "Marine debris along the Caribbean coast of Panama". In: *Marine Pollution Bulletin* 26.6 (1993), pp. 317–324. ISSN: 0025-326X. DOI: https://doi.org/10.1016/0025-326X(93)90574-4. URL: http://www.sciencedirect.com/science/article/pii/0025326X93905744.

[22] Stefanie Reinold. "Plastic pollution on eight beaches of Tenerife (Canary Islands, Spain): An annual study: Wind and Wave parameters". In: (Mar. 2020). DOI: 10.17632/f7ntbw4rt6.1. URL: https://mendeley.figshare.com/articles/Plastic_pollution_on_eight_beaches_of_Tenerife_Canary_Islands_Spain_An_annual_study_Wind_and_Wave_parameters/11972994.

[23] Anthony Cheshire et al. *UNEP/IOC Guidelines on Survey and Monitoring of Marine Litter*. Jan. 2009.

[24] Ronen Alkalay, Galia Pasternak, and Alon Zask. "Clean-coast index—A new approach for beach cleanliness assessment". In: *Ocean and Coastal Management* 50 (Dec. 2007). DOI: 10.1016/j.ocecoaman.2006.10.002.

[25] Michelle Portman and Ruth Brennan. "Marine litter from beach-based sources: Case study of an Eastern Mediterranean coastal town". In: *Waste Management* 69 (Aug. 2017). DOI: 10.1016/j.wasman.2017.07.040.

[26] Arun Kumar A. et al. "Preliminary study on marine debris pollution along Marina beach, Chennai, India". In: *Regional Studies in Marine Science* 5 (2016), pp. 35–40. ISSN: 2352-4855. DOI: https://doi.org/10.1016/j.rsma.2016.01.002. URL: http://www.sciencedirect.com/science/article/pii/S2352485516300020.

[27] Allan Williams et al. "Distribution of beach litter along the coastline of Cádiz, Spain". In: *Marine pollution bulletin* 107 (Apr. 2016). DOI: 10.1016/j.marpolbul.2016.04.015.

[28] Hannah Earp and Arianna Liconti. *Science for the future: Citizen science in marine research and conservation*. Feb. 2019. DOI: 10.13140/RG.2.2.34382.92483.

[29] Helen Roy et al. "Understanding citizen science and environmental monitoring". In: *Final Report on Behalf of UK-EOF* (Jan. 2012). DOI: 10.1002/9781118360989.ch6.

[30] Graham Forrester et al. "Comparing monitoring data collected by volunteers and professionals shows that citizen scientists can detect long-term change on coral reefs". In: *Journal for Nature Conservation* 24 (Apr. 2015). DOI: 10.1016/j.jnc.2015.01.002.

[31] Macarena Bravo et al. “Anthropogenic debris on beaches in the SE Pacific (Chile): Results from a national survey supported by volunteers”. In: *Marine pollution bulletin* 58 (Sept. 2009), pp. 1718–26. DOI: 10.1016/j.marpolbul.2009.06.017.

[32] Henry Carson et al. “Tracking the sources and sinks of local marine debris in Hawai’i. Marine Environmental Research, 84, 76-83”. In: *Marine environmental research* 84 (Dec. 2012). DOI: 10.1016/j.marenvres.2012.12.002.

[33] Valeria Hidalgo-Ruz and Martin Thiel. “The Contribution of Citizen Scientists to the Monitoring of Marine Litter”. In: Jan. 2017, p. 125. ISBN: 9780128122716. DOI: 10.1016/B978-0-12-812271-6.00123-X.

[34] Anne Bauer-Civiello, Jennifer Loder, and Mark Hamann. “Using citizen science data to assess the difference in marine debris loads on reefs in Queensland, Australia”. In: *Marine Pollution Bulletin* 135 (Oct. 2018), pp. 458–465. DOI: 10.1016/j.marpolbul.2018.07.040.

[35] Marina Locritani, Silvia Merlino, and Marinella Abbate. “Assessing the citizen science approach as tool to increase awareness on the marine litter problem”. In: *Marine Pollution Bulletin* 140 (Mar. 2019), pp. 320–329. DOI: 10.1016/j.marpolbul.2019.01.023.

[36] J. R. Jambeck and K. Johnsen. “Citizen-Based Litter and Marine Debris Data Collection and Mapping”. In: *Computing in Science Engineering* 17.4 (2015), pp. 20–26.

[37] Mark Anthony Browne et al. “Spatial and Temporal Patterns of Stranded Intertidal Marine Debris: Is There a Picture of Global Change?” In: *Environmental Science and Technology* 49 (May 2015). DOI: 10.1021/es5060572.

```
#```{r}
dataModel <- list.files(path = "data/debris/", full.names = TRUE) %>%
  lapply(FUN = read_csv, col_types = "ififiddddcfcif") %>%
  reduce(rbind)
dataModel$Time <- data$Timestamp %>%
  parse_datetime(format = "%Y%m%d%H%M%S")
dataModel$Timestamp <- NULL
dataModel <- dataModel %>% select(-ListID,-ListName)
plastic_ordered <- dataModel %>%
  filter(`Material Description` == "PLASTIC") %>%
  select(ItemName, Quantity) %>%
  group_by(ItemName) %>%
  summarise(Total = sum(Quantity)) %>%
  arrange(desc(Total))
dataModel$ItemName <- factor(dataModel$ItemName, plastic_ordered$ItemName)
recategorise2 <- function(x){
  out = ""
  if(x %in% c(1,4,6,22)){out = "Cigarette related waste"}
  if(x %in% c(2,3,7,9,10,17,23,11)) out = "Food related waste"
  if(x %in% c(8,14,15,16,18,19,21,20)) out = "Other"
  if(x %in% c(12,13)) out = "Plastic bags and Styrofoam packaging"
  if(x %in% c(5,23,24,25)) out = "Fragments"
  if(out == "") stop(paste("Error in recategorise:", x))
  return(out)
}
plastic_typesModel <- dataModel %>%
  filter(`Material Description` == "PLASTIC") %>%
  select(ItemName, ItemID) %>%
  distinct() %>%
```

```
  mutate(label = 1:n()) %>%
  mutate(category = purrr::map(label, recategorise2)) %>%
  mutate(category = as_factor(as.character(category))) %>%
  select(ItemID, category)
plasticModel <- dataModel %>%
  filter(`Material Description` == "PLASTIC") %>%
  full_join(plastic_types, by = "ItemID")
plasticN <- plasticModel %>%
  mutate(month = month(Time, label = FALSE),
         year = as.integer(year(Time))) %>%
  filter(year > 2010) %>%
  group_by(month, year, category) %>%
  summarise(`Total Quantity` = sum(Quantity)) %>%
  ggplot(aes(x = month, y = `Total Quantity`, fill = category)) +
    geom_col(colour = "black", size = 0.2, position = "fill") +
    facet_wrap(~year, nrow = 2) +
    scale_fill_viridis(discrete = TRUE, option = "plasma") +
    xlab("Month") +
    ylab("Proportion of Items") +
    ggtitle("Rel. frequencies of observed plastic waste by category") +
    scale_x_continuous(breaks = 1:12) +
    theme(panel.grid.major.x = element_blank(),
          panel.grid.minor.x = element_blank()) +
    guides(fill=guide_legend(title="Category"))
#```

#```{r}
library(dplyr)
df12N <- plasticN  %>%
  filter(year == 2012) %>%
  group_by(year, month) %>%
  mutate(freq = `Total Quantity` / sum(`Total Quantity`))
df13N <- plasticN  %>%
  filter(year == 2013) %>%
  group_by(year, month) %>%
  mutate(freq = `Total Quantity` / sum(`Total Quantity`))
df14N <- plasticN  %>%
  filter(year == 2014) %>%
  group_by(year, month) %>%
  mutate(freq = `Total Quantity` / sum(`Total Quantity`))
df15N <- plasticN  %>%
  filter(year == 2015) %>%
  group_by(year, month) %>%
  mutate(freq = `Total Quantity` / sum(`Total Quantity`))
df16N <- plasticN  %>%
  filter(year == 2016) %>%
  group_by(year, month) %>%
  mutate(freq = `Total Quantity` / sum(`Total Quantity`))
df17N <- plasticN  %>%
  filter(year == 2017) %>%
  group_by(year, month) %>%
  mutate(freq = `Total Quantity` / sum(`Total Quantity`))
df18N <- plasticN  %>%
```

```
  filter(year == 2018) %>%
  group_by(year, month) %>%
  mutate(freq = `Total Quantity` / sum(`Total Quantity`))
df19N <- plasticN  %>%
  filter(year == 2019) %>%
  group_by(year, month) %>%
  mutate(freq = `Total Quantity` / sum(`Total Quantity`))
dfTotN <- rbind(df12N, df13N, df14N, df15N, df16N, df17N, df18N, df19N)
# plot for observing the data
(time_plotfr2N <- ggplot(dfTotN, aes(x = year, y = freq, color=category, fill = category)) +
  geom_smooth(method="lm", level=0.95) +
  theme_bw() +
  xlab("Years") +
  ylab("relative frequency") +
  ggtitle("portion of plastic") +
  expand_limits(y=0) +
  scale_y_continuous() +
  scale_x_continuous()+
  theme(legend.position="bottom")+
  theme(legend.text = element_text(size=5, face="bold")))
#```



#We see here a graphical representation of the relative frequency of the 5 different categories of plas

#```{r}
# create train and test set
n <- nrow(dfTotN)  # Number of observations
ntrain <- round(n*0.75)  # 75% for training set
set.seed(314)    # Set seed for reproducible results
tindex <- sample(n, ntrain)   # Create a random index
train_dfTotN <- dfTotN[tindex,]   # Create training set
test_dfTotN <- dfTotN[-tindex,]
# modelling for category "Cigarette related waste"
train_Cigrel <- train_dfTotN %>%
  filter(category=="Cigarette related waste") %>%
  group_by(year)
test_Cigrel <- test_dfTotN %>%
  filter(category=="Cigarette related waste") %>%
  group_by(year)
set.seed(1234)
train_Cigrel.modelN <- lm(freq ˜ year, data = train_Cigrel)
summary(train_Cigrel.modelN)
print("PREDICTION")
pred_Cigrel <- predict(train_Cigrel.modelN, test_Cigrel)
summary(pred_Cigrel)
actuals_predsCigrel <- data.frame(cbind(actuals=test_Cigrel$freq, predicteds=pred_Cigrel))
head(actuals_predsCigrel)
correlation_accuracy <- cor(actuals_predsCigrel)
min_max_accuracy <- mean(apply(actuals_predsCigrel, 1, min) / apply(actuals_predsCigrel, 1, max))
correlation_accuracy
min_max_accuracy
```

```
#```


#```{r}
# modelling for category "Food related waste"
train_Foodrel <- train_dfTotN %>%
  filter(category=="Food related waste") %>%
  group_by(year)
test_Foodrel <- test_dfTotN %>%
  filter(category=="Food related waste") %>%
  group_by(year)
set.seed(1234)
train_Foodrel.modelN <- lm(freq ~ year, data = train_Foodrel)
summary(train_Foodrel.modelN)
print("PREDICTION")
pred_Foodrel <- predict(train_Foodrel.modelN, test_Foodrel)
summary(pred_Foodrel)
actuals_predsFoodrel <- data.frame(cbind(actuals=test_Foodrel$freq, predicteds=pred_Foodrel))
head(actuals_predsFoodrel)
correlation_accuracy <- cor(actuals_predsFoodrel)
min_max_accuracy <- mean(apply(actuals_predsFoodrel, 1, min) / apply(actuals_predsFoodrel, 1, max))
correlation_accuracy
min_max_accuracy
#```


#```{r}
# modelling for category "Fragments"
train_Frag <- train_dfTotN %>%
  filter(category=="Fragments") %>%
  group_by(year)
test_Frag <- test_dfTotN %>%
  filter(category=="Fragments") %>%
  group_by(year)
set.seed(1234)
train_Frag.modelN <- lm(freq ~ year, data = train_Frag)
summary(train_Frag.modelN)
print("PREDICTION")
pred_Frag <- predict(train_Frag.modelN, test_Frag)
summary(pred_Frag)
actuals_predsFrag <- data.frame(cbind(actuals=test_Frag$freq, predicteds=pred_Frag))
head(actuals_predsFrag)
correlation_accuracy <- cor(actuals_predsFrag)
min_max_accuracy <- mean(apply(actuals_predsFrag, 1, min) / apply(actuals_predsFrag, 1, max))
correlation_accuracy
min_max_accuracy
#```


#```{r}
# modelling for category "Other"
train_Other <- train_dfTotN %>%
  filter(category=="Other") %>%
```

```
  group_by(year)
test_Other <- test_dfTotN %>%
  filter(category=="Other") %>%
  group_by(year)
set.seed(1234)
train_Other.modelN <- lm(freq ~ year, data = train_Other)
summary(train_Other.modelN)
print("PREDICTION")
pred_Other <- predict(train_Other.modelN, test_Other)
summary(pred_Other)
actuals_predsOther <- data.frame(cbind(actuals=test_Other$freq, predicteds=pred_Other))
head(actuals_predsOther)
correlation_accuracy <- cor(actuals_predsOther)  # 5.31%
min_max_accuracy <- mean(apply(actuals_predsOther, 1, min) / apply(actuals_predsOther, 1, max))
correlation_accuracy
min_max_accuracy
#```


#```{r}
# modelling for category "Plastic bags and Styrofoam packaging"
train_Plbag <- train_dfTotN %>%
  filter(category=="Plastic bags and Styrofoam packaging") %>%
  group_by(year)
test_Plbag <- test_dfTotN %>%
  filter(category=="Plastic bags and Styrofoam packaging") %>%
  group_by(year)
set.seed(1234)
train_Plbag.modelN <- lm(freq ~ year, data = train_Plbag)
summary(train_Plbag.modelN)
print("PREDICTION")
pred_Plbag <- predict(train_Plbag.modelN, test_Plbag)
summary(pred_Plbag)
actuals_predsPlbag <- data.frame(cbind(actuals=test_Plbag$freq, predicteds=pred_Plbag))
head(actuals_predsPlbag)
correlation_accuracy <- cor(actuals_predsPlbag)
min_max_accuracy <- mean(apply(actuals_predsPlbag, 1, min) / apply(actuals_predsPlbag, 1, max))
correlation_accuracy
min_max_accuracy
#```

#```{r}
plastic_category <-c("cigarette related waste", "food related waste","Fragments",
                      "Other","Plastic bags and Styrofoam packaging" )
slope_scores <- c(-0.063,0.038, 0.034, 0.000, 0.001)
slope_interpretation <-c("downward", "upward", "upward", "steady", "steady")
p_value<-c("<0.05","<0.05","<0.05", ">0.05",">0.05")
adjRsquared <- c(0.4416, 0.4324, 0.2177,  -0.0146, -0.01195)
corr_accuracy<-c(0.83, 0.50, 0.41, -0.36,-0.12)
min_max_Acc<-c(0.78,0.77,0.69, 0.64 ,0.61)
score_table1 <- data.frame(plastic_category, p_value,slope_scores, slope_interpretation, adjRsquared)
score_table2 <- data.frame(plastic_category, corr_accuracy, min_max_Acc)
score_table1
```

```
score_table2
```
```

## Error:  attempt to use zero-length variable name
```