explore_outliers_OCdata

Corey Clatterbuck

Purpose

Identify potential outliers in ocean conservancy data that may be due to data entry issues. The identified issues (outlined in issue #5) so far include:

1. Large numbers (>1000) of `People` and `Adults` for some cleanups
2. Some cleanups include just a single trash item or category, including those with large numbers of people

I wanted to visualize how much data is potentially impacted by these issues to better determine how much data to drop.

Load data & libraries

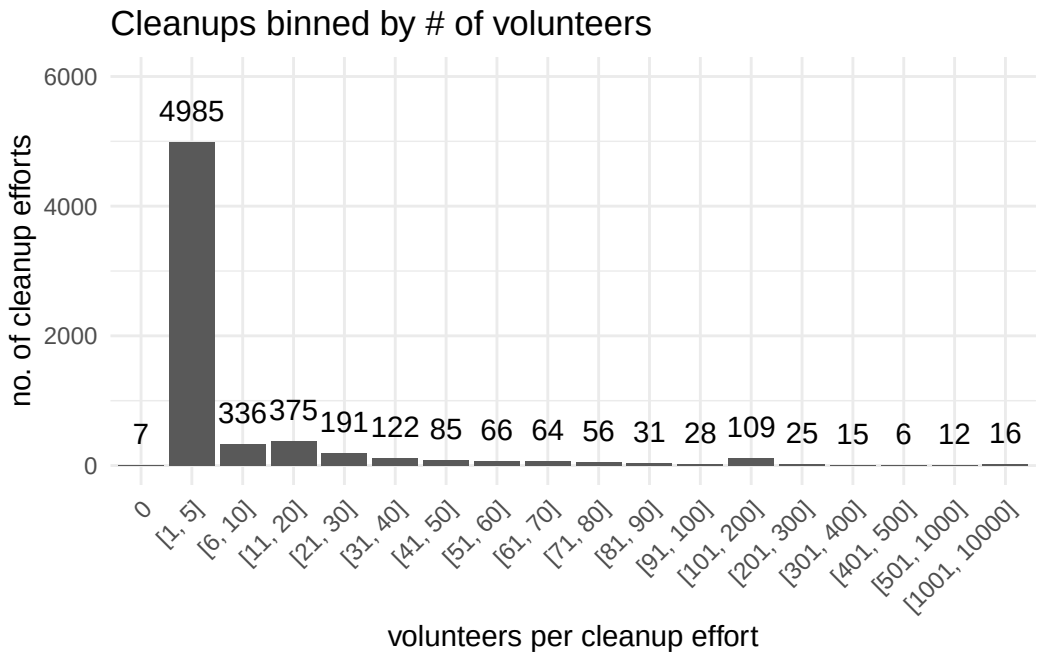
Code for accomplishing this is available in the .qmd file.

OC data cleanup

Here I reduce OC data to the dates of Coastal Cleanup Day & remove the columns (plastic types) that we do not use – these are columns that do not include any plastic items or 'pieces' that are not identifiable to a plastic category. The rows are then filtered to remove any cleanups lacking counts for all plastic categories. This last step removes ~1600 cleanups.

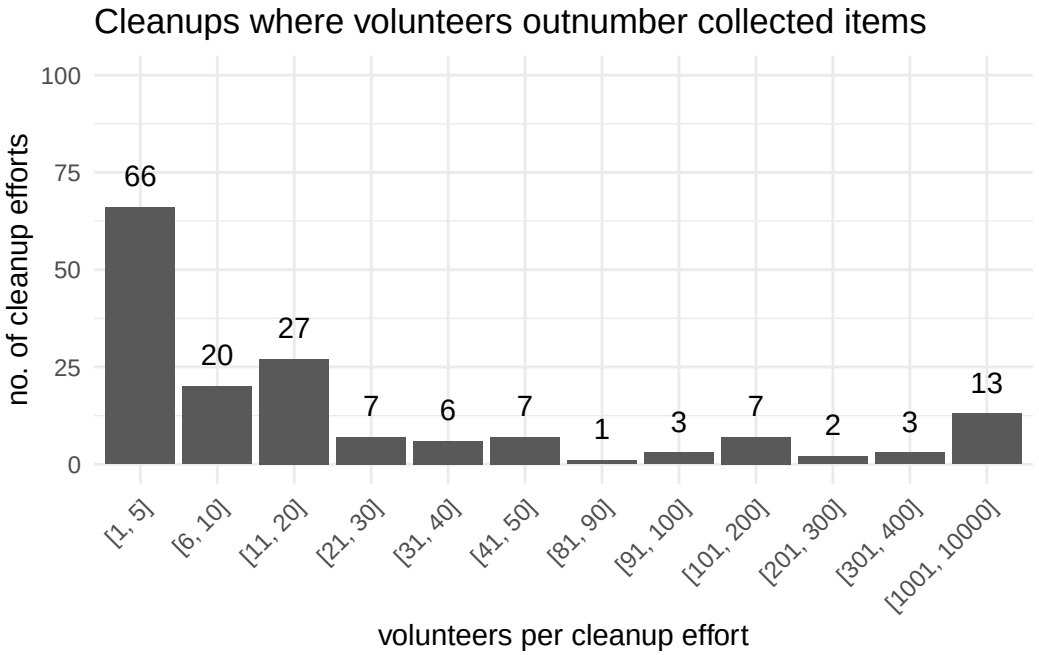
The code for accomplishing this is in the .qmd file

Bin cleanups by people



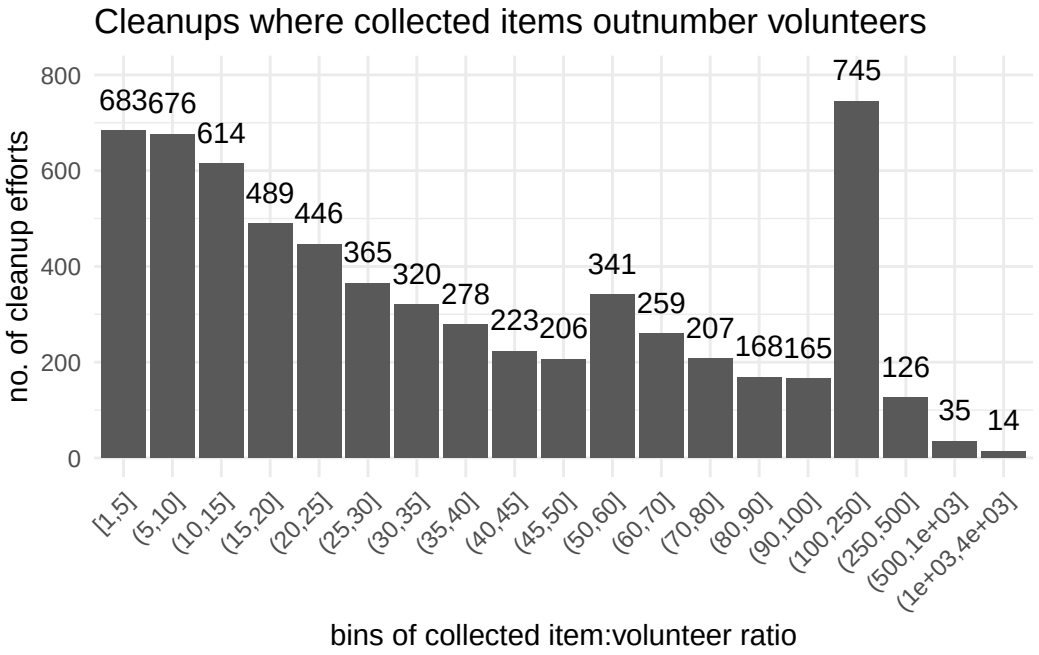
How many cleanups have more people than collected items?

And potentially, what bin do these groups fall in?



While most of these cleanups were in groups of few people, note that 13 of 16 cleanups occurring in the [1001, 10000] bin also had a item:people ratio < 1. All item:people ratios less than 1 total 163 cleanups, which is ~2.5% of the data filtered to this point.

For reference, here is a plot of the collected item:people ratios for the remaining data. I remove cleanups where people $= 0$ in the code chunk as well because `cut` doesn't play nicely with infinite numbers. Note the binned intervals change on the x-axis as we move from left to right.

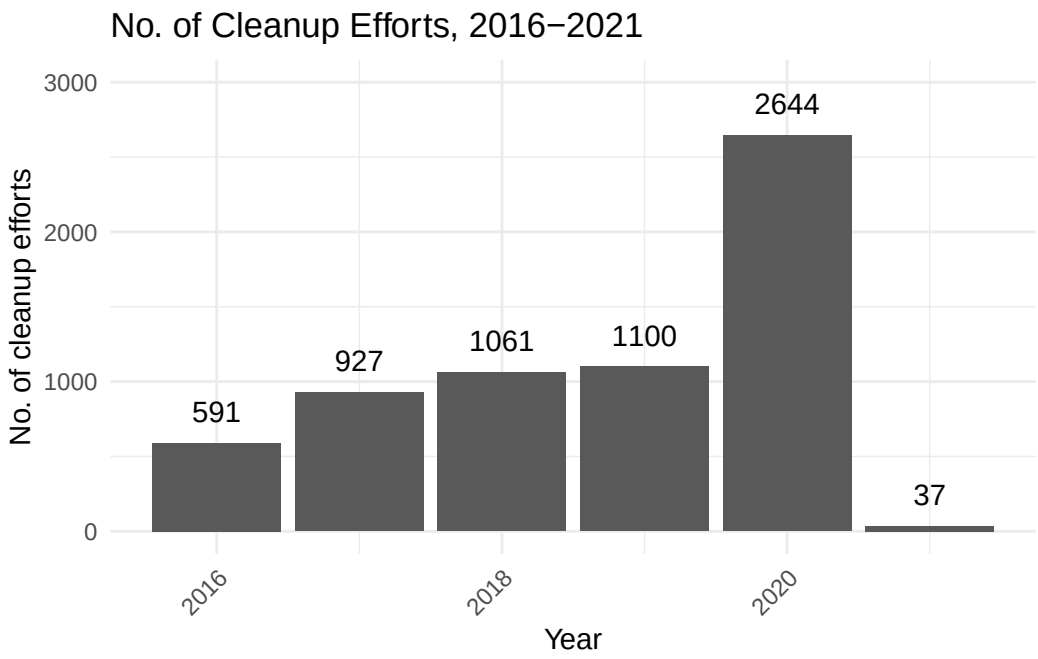


Summary

After filtering the OC data to Coastal Cleanup Day dates, we have **8143** cleanups. This was completed in previous scripts but wanted to include here as well.

- Remove unidentifiable plastic categories & reduce to cleanups with identifiable items only: -1614, **6529**
- Remove cleanups where volunteers outnumber collected items: -162, **6367**
- Remove cleanups with 0 people, as you cannot account for effort: -7, **6360 cleanups**

Here's a breakdown of the number of cleanups by year – note that there are many more cleanups for 2020 as the whole month of September was labeled "Coastal Cleanup Month", rather than a single day, due to the COVID pandemic. Also, 2021's data is incomplete. It would be good to get an updated dataset for 2021-2022 if easy for OC folks to accomplish.



Code to clean raw OC data

The code below can be used to load & clean the ocean conservancy data from the raw file:

```
library(tidyverse)
library(lubridate)
library(readxl)
library(here)
library(janitor)

oc_raw <- read_excel(here("data", "OceanConservancy_CA.xlsx"),
                     col_types = c("numeric", "text", "text", "text", "text",
                                   "text", "date", "text", "numeric", "numeric",
                                   "numeric", "numeric", "numeric", "numeric", "numeric",
                                   "numeric", "numeric", "numeric", "numeric", "numeric",
                                   "numeric", "numeric", "numeric", "numeric", "numeric",
                                   "numeric", "numeric", "numeric", "numeric", "numeric",
                                   "numeric", "numeric", "numeric", "numeric", "numeric",
                                   "numeric", "numeric", "numeric", "numeric", "numeric",
                                   "numeric", "numeric", "numeric", "numeric", "numeric",
```

```
                                    "numeric", "numeric", "numeric", "numeric", "numeric",
                                    "numeric", "numeric", "numeric", "numeric", "numeric",
                                    "numeric", "numeric", "numeric", "numeric", "numeric"))

## reduce dates
ccd_dates <- c("2016-09-17", "2017-09-16", "2018-09-15", "2019-09-21", "2021-09-19")
ccd_dates <- as.Date(ccd_dates)
oc_use <- oc_raw %>%
  mutate(`Cleanup Date` = as.Date(`Cleanup Date`)) %>%
  dplyr::filter(`Cleanup Date` %in% ccd_dates |
                  (`Cleanup Date` >= as.Date('2020-09-01') & `Cleanup Date` <= as.Date('20
levels(as.factor(oc_use$`Cleanup Date`)) ## double check
```

```
 [1] "2016-09-17" "2017-09-16" "2018-09-15" "2019-09-21" "2020-09-01"
 [6] "2020-09-02" "2020-09-03" "2020-09-04" "2020-09-05" "2020-09-06"
[11] "2020-09-07" "2020-09-08" "2020-09-09" "2020-09-10" "2020-09-11"
[16] "2020-09-12" "2020-09-13" "2020-09-14" "2020-09-15" "2020-09-16"
[21] "2020-09-17" "2020-09-18" "2020-09-19" "2020-09-20" "2020-09-21"
[26] "2020-09-22" "2020-09-23" "2020-09-24" "2020-09-25" "2020-09-26"
[31] "2020-09-27" "2020-09-28" "2020-09-29" "2020-09-30" "2021-09-19"
```

```
## remove unidentifiable pieces
oc_use <- oc_use |>
  dplyr::rename("Fishing Nets" = "Fishing Net & Pieces") |>
  dplyr::select(-ends_with(c("(Clean Swell)", "Pieces", "Collected"))) |>
  dplyr::rename("Fishing Net & Pieces" = "Fishing Nets") |>
  dplyr::filter(!if_all(15:57, is.na))  ## removes ~1600 cleanups

## remove collected item:people ratio less than one, as well as 0s. Remove cols use to cal
oc_use <- oc_use |>
  dplyr::mutate(total_collected = rowSums(across(15:57), na.rm = TRUE),
                ratio_collected = round(total_collected/People, 4)) |>
  relocate(ratio_collected, .after = People) |>
  dplyr::filter(ratio_collected > 0.9999) |>
  dplyr::filter(is.finite(ratio_collected)) |>
  dplyr::select(-total_collected, -ratio_collected)
```