Processing Data for CapOnLitter

A guide written by Jody Holland

The CapOnLitter data is supplied in the form of an .xlxs file wherein each page is a different month. The dataset is written in Spanish. The method of reporting the data has shifted over the past 3 years, such as in terms of whether the numbers of items found is the raw data, or if it is presented as the average over a 100m and 1000m space. Other notable changes are the merging of litter types into one for ease of sampling, the removal of temperature data, the inclusion of tide height and wind direction data by 2023, and a change in the way the spreadsheet pages are formatted.

To better understand the similarities and differences within the evolution of the way the data has been organised, Table 1.1 below shows the change from March 2021 to March 2023:

Feature	March 2021	March 2023
Wind Direction	Not Present	Shown as an abbreviated Cardinal Direction (E.G. NW)
Wind Intensity/Speed	Shown as a km/h value in the climate conditions cell	Shown as a number value in the same cell as the Wind Direction
Humidity	Shown as a percentage in the climate conditions cell	Not Present
Temperature	Shown as a value in celsius in the climate conditions cell	Not Present
Animals present	Shown as a YES or NO value	Shown as a 1 or 0 value
Start Coords	Shown as an x and y value in a specific cell	Shown as the first two values in a coords cell
End Coords	Shown as an x and y value in a specific cell	Shown as the last two values in a coords cell, after the start coords
Zone	Provided	Not Present
Litter Type ID	Provided	Not Present
Data on Textiles	Provided and broken down in subcategories	Textile related litter such as clothing is put under "Others"
Aggregated to 100m average	Yes	No
Aggregated to 1000m average	Yes	No

Goals:

- Data between months to be comparable
- Geodata to be preserved
- IDs on litter typology to be present
- Formatted as geodata files such as .shp and .geojson

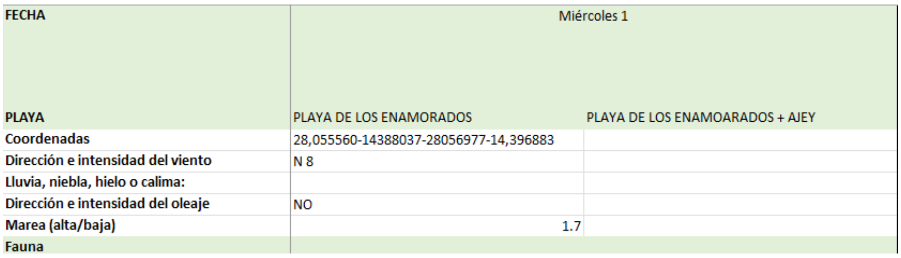
Making the months align

This is probably the hardest task. The first step is understanding the differences between the months and deciding how to aggregate the more detailed early months so that they align with the later months. For instance, say shoes and clothes were separate categories in the 2021, and by 2023 they are grouped together, then one would need to add the columns for shoes and clothes in 2021 into one column so that it aligns.This is likely going to be a difficult process that requires a lot of attention to ensure that the right columns are aggregated together and thus the data is able to be used in accurate comparative analysis.

Another question is the creation of the comparable index wherein the data is aggregated by 100m or 1000m averages. I think this is a good system and allows for beaches of different sizes to be compared. My suggestion is that all the data is formatted as 100m averages and the length of the sampling transect is also provided. This means that all that is required to reserve the aggregation is a simple formula as such:

$$Raw\ Litter\ Data\ =\ \frac{Aggregated\ Data * Length\ of\ Beach}{100}$$

There is also the hard task of disentangling the excel data. Currently the data requires pivoting, this means that the columns become rows and the rows become columns. Thus each column shifts from being a beach to being a variable. This is important because it allows us to import the data into R, for both analysis and file conversion. In practising on the March 2021 and March 2023 data this task is hard, taking between 3-4 hours per dataset. One of the trickiest parts is understanding how to deal with merged columns and inferred data. For example, in this image you see that both these beach samples share the same date:



FECHA	Miércoles 1	
PLAYA	PLAYA DE LOS ENAMORADOS	PLAYA DE LOS ENAMOARADOS + AJEY
Coordenadas	28,055560-14388037-28056977-14,396883	
Dirección e intensidad del viento	N 8	
Lluvia, niebla, hielo o calima:		
Dirección e intensidad del oleaje	NO	
Marea (alta/baja)	1.7	
Fauna		

However for “PLAYA DE LOS ENAMOARADOS + AJEY'' there is no data given on sampling coordinates or climate data. Obviously the coordinates are different from “PLAYA DE LOS ENAMORADOS”, but the inference is that they share the data on wind direction, speed, and tidal information. Thus, data in theory needs to be copied over in every similar situation from its left hand neighbour, this is a task that is quite time consuming, though I am working on it some R script can speed up the process.

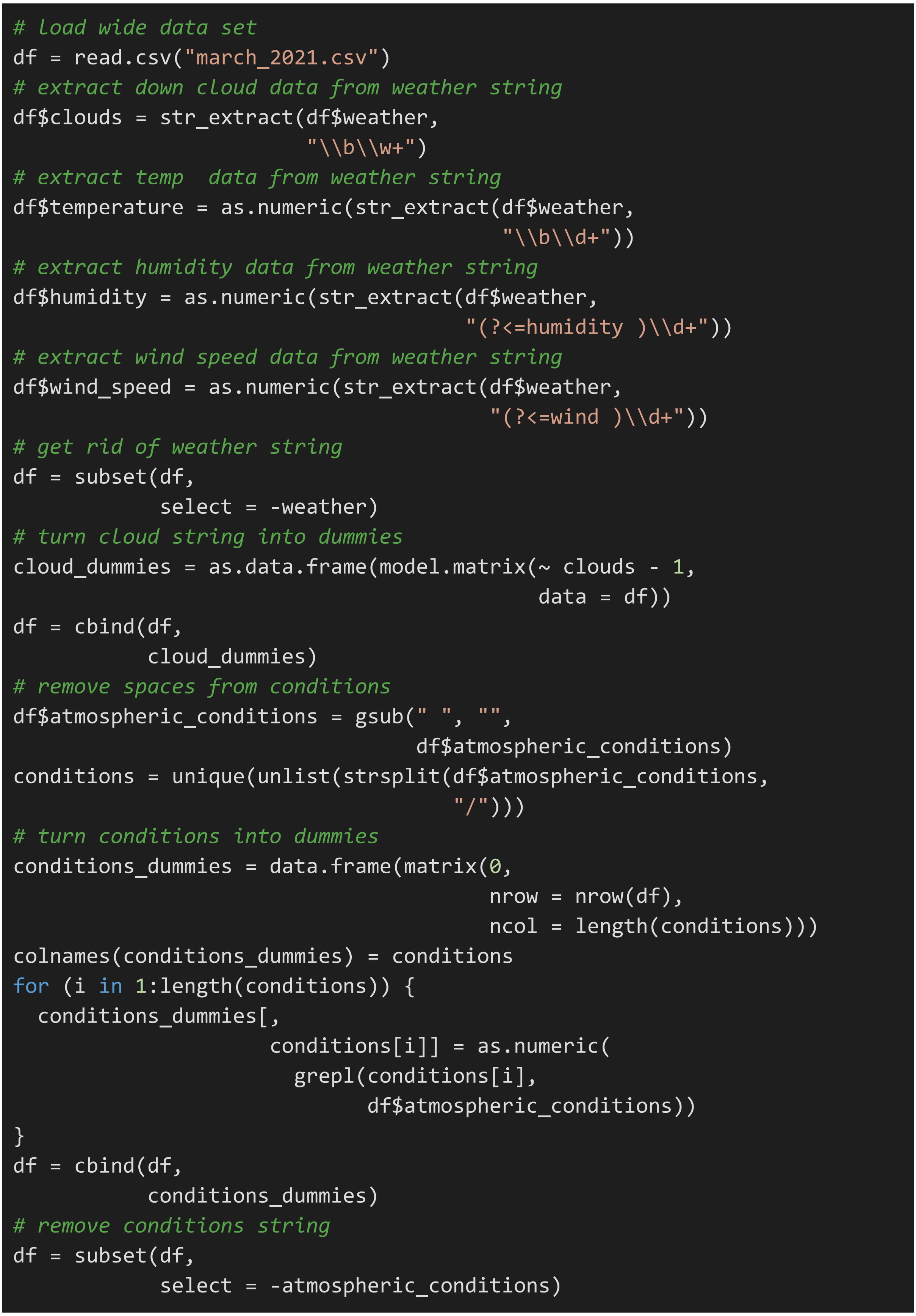
Creating Dummy Variables

The raw data contains a lot of categorical information. This is often in the form of lines of text written by surveyors regarding the conditions on the beaches on the day that they sampled. To transform this categorical data into usable numerical values for analysis, one option is to break the text down into dummies. A dummy variable in quantitative analysis simply means a binary indicator of either yes or no, as shown by a 1 or 0 relating to a certain condition.

For example, in a description of the meteorological conditions on the beach, it may be that the samplers described a fair level of cloud cover and a moderately strong wind. Breaking this down into dummy variables would mean that across a series of individual variables indicating different levels of cloud cover or wind speed for most of the beach would have a value of 0 but for the category of fair cloud cover and moderate wind speed, the beach would receive a value of 1. It is my conviction that for statistically modelling it may be critical that the data is transformed in such a way. To summarise my reasoning here:

- Categorical data can be transformed into numerical values through dummy variables.
- Dummy variables are binary indicators that represent either yes or no for certain conditions.
- Breaking down categorical data into dummy variables can help make it more usable for statistical analysis.

In my opinion, the best way to transform text into this numerical categorical dummy variable form is to use the tools in R for handling string data. This entails breaking the string down into its constituent parts based on which categorical variable is being referred to and then forming dummy variables from this data. On the following page, you can see the code I used to do this for the March 2021 data.



```
# load wide data set
df = read.csv("march_2021.csv")
# extract down cloud data from weather string
df$clouds = str_extract(df$weather,
                        "\\b\\w+")
# extract temp  data from weather string
df$temperature = as.numeric(str_extract(df$weather,
                                        "\\b\\d+"))
# extract humidity data from weather string
df$humidity = as.numeric(str_extract(df$weather,
                                     "(?<=humidity )\\d+"))
# extract wind speed data from weather string
df$wind_speed = as.numeric(str_extract(df$weather,
                                       "(?<=wind )\\d+"))
# get rid of weather string
df = subset(df,
            select = -weather)
# turn cloud string into dummies
cloud_dummies = as.data.frame(model.matrix(~ clouds - 1,
                                           data = df))
df = cbind(df,
           cloud_dummies)
# remove spaces from conditions
df$atmospheric_conditions = gsub(" ", "",
                                 df$atmospheric_conditions)
conditions = unique(unlist(strsplit(df$atmospheric_conditions,
                                    "/")))
# turn conditions into dummies
conditions_dummies = data.frame(matrix(0,
                                       nrow = nrow(df),
                                       ncol = length(conditions)))
colnames(conditions_dummies) = conditions
for (i in 1:length(conditions)) {
  conditions_dummies[,
                     conditions[i]] = as.numeric(
                       grepl(conditions[i],
                             df$atmospheric_conditions))
}
df = cbind(df,
           conditions_dummies)
# remove conditions string
df = subset(df,
            select = -atmospheric_conditions)
```

Geodata

Each beach in theory has a start and end coordinate of the sample. This presents us with several options in deciding how to format the data as a geoobject. Initially I have been calculating the midpoint of the transect and using that as a point data. However, it is likely possible to format the data instead as lines/vectors. This is in a similar fashion to roadway datasets. This should be possible using the "sf" package for R. It may also be worth keeping the midpoint coords in the data, in case someone wishes to use the data as point daThista. In terms of which file formats to use, I suggest that we use both .shp and .geojson. should keep file sizes down, and make sure that the data is able to open in major GIS and data analysis services such as ArcGIS and R.

General steps for processing the data into geodata format

1. When loading into Excel the first task is to remove all of the previously merged cells so that each sample area and day has its own respective information regarding the meteorological conditions and the area of sampling.
2. Following this step, the second task is to remove all of the unnecessary data and formatting.
3. Next each individual month needs to be saved as a CSV file.
4. The file can then be loaded into an R workspace
5. Using the tidy verse family of commands the table can now be pivoted from long to wide format. in this way of compiling the data each row represents an individual sample area on a given day.
6. Using the string manipulation commands in R the meteorological data can be broken down into dummies.
7. Again using the string manipulation tools the coordinate data can be extracted.
8. To determine the midpoint of the sampling a shapefile of the coastline of Fuerteventura must be loaded.
9. loading both the start and endpoint onto this shape file, the midpoint of the sample area can be determined along the coast as well as they length of the sample area
10. This midpoint can then be set as the geompoint information for a shapefile formating of the data
11. Furthermore using the coastline shapefile the length of the sample area can be determined
12. Using this length the rate of little collection for 100 and 1,000 metres can also be determined
13. It may also be prudent to develop alternative formatting of the data in the shape files not related to points but rather to vector lines in the area samples itself.