Seabirds\_project\_description

Neil Plenderleith

2022-07-31

# Seabirds Project Neil Plenderleith CdoeClan DR16

## Brief Introduction to the Dataset

The dataset is in .xlx format with 4 sheets. The main two are bird sightings and the ship they are sighted from. There are also 2 sheets which describe the column data and abbreviations etc. There are a lot of NA values! I can understand a lot of these wil be from incomplete sightings where not all data points can possible be collected from every sighting.

## List of Assumptions

I have assumed the cleaned data will go on to be used in other analyses and therefore I have not trimmed it massively. Were this not the case I would have cut 45 columns or so which were not required for our 5 tasks. I did trim two very nearly empty columns in the end (4 more were borderline) but I have erred on the side of keeping data in however incomplete rather than dropping columns at will.

## Steps to Clean Data

As mentioned in assumptions I probably erred on the side of leaving data in rather than dropping it. Otherwise I cleaned the column names, changed a couple of very long names and dropped two 99.99% empty columns.

library(readxl)  
library(here)

## here() starts at C:/Users/neilp/Documents/CODECLAN/dirty\_data\_codeclan\_project\_neil\_plenderleith/task\_3

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4  
## ✔ tibble 3.1.7 ✔ dplyr 1.0.9  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.0  
## ✔ readr 2.1.2 ✔ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(janitor)

##   
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':  
##   
## chisq.test, fisher.test

birds <- read\_xls(here("raw\_data/seabirds.xls"),   
 sheet = "Bird data by record ID")

## Warning: Expecting logical in I21756 / R21756C9: got 'M'

ships <- read\_xls(here("raw\_data/seabirds.xls"),   
 sheet = "Ship data by record ID")  
  
  
seabirds <- left\_join(birds, ships, "RECORD ID")  
  
# so from the notes and from the above investigation we can:  
  
# 1 sort column names  
# 2 drop 2 columns with only 1, 4 values out of 49019 (4 more cols were   
# borderline but we will keep these for now)  
  
# lets sort all these uppercase column names  
seabirds <- clean\_names(seabirds)  
  
# lets drop cols sex and sal. sex only has 1/49019 values, SAL has 3/49017  
seabirds <- seabirds %>%   
 select(-sex, -sal)  
  
# lets change some long column names  
seabirds <- seabirds %>%   
rename(species\_common\_name = species\_common\_name\_taxon\_age\_sex\_plumage\_phase,  
 species\_scientific\_name = species\_scientific\_name\_taxon\_age\_sex\_plumage\_phase)  
  
# Can we use this data now? It has a lot of NAs but i suppose a lot are just  
# the way the data is recorded and the way the dataset is setup. Not every  
# observation would have all fields completed.  
  
seabirds %>%   
write\_csv(here("clean\_data/seabirds.csv"))

## Tasks

### Task 1 Which bird had the most individual sightings?

# group by bird name and take a count of occurances, join this back into the table  
# to allow more names info to be pulled, arrange and slice head to get top hit  
seabirds %>%   
 group\_by(species\_common\_name) %>%   
 summarise(count\_of\_sightings = n()) %>%   
 left\_join(seabirds, "species\_common\_name") %>%   
 select(species\_common\_name,   
 species\_scientific\_name,   
 species\_abbreviation, count\_of\_sightings) %>%  
 arrange(desc(count\_of\_sightings)) %>%   
 slice\_head()

## # A tibble: 1 × 4  
## species\_common\_name species\_scienti… species\_abbrevi… count\_of\_sighti…  
## <chr> <chr> <chr> <int>  
## 1 Wandering albatross sensu … Diomedea antipo… DIOANTEXU PL2 3104

# Answer: Wandering albatross sensu lato PL2, count 3104

### Task 2 Which bird had the highest total count?

# group by bird name and take a sum of count, join this back into the table  
# to allow more names info to be pulled, arrange and slice head to get top hit  
seabirds %>%   
 group\_by(species\_common\_name) %>%   
 summarise(total\_count = sum(count)) %>%   
 left\_join(seabirds, "species\_common\_name") %>%   
 select(species\_common\_name,   
 species\_scientific\_name,   
 species\_abbreviation, total\_count) %>%  
 arrange(desc(total\_count)) %>%   
 slice\_head()

## # A tibble: 1 × 4  
## species\_common\_name species\_scientific\_name species\_abbreviation total\_count  
## <chr> <chr> <chr> <dbl>  
## 1 Black-billed gull Larus bulleri LARBUL 7732

# Answer: A Black-billed gull, count = 7732

### Task 3 Which bird had the highest total count above a latitude of -30?

# group by bird name, filter by lat >-30 and take a summary of the count column for each  
seabirds %>%   
 filter(lat > -30) %>%   
 group\_by(species\_common\_name) %>%   
 summarise(total\_count = sum(count)) %>%   
 left\_join(seabirds, "species\_common\_name") %>%   
 select(species\_common\_name,   
 species\_scientific\_name,   
 species\_abbreviation, total\_count) %>%  
 arrange(desc(total\_count)) %>%   
 slice\_head()

## # A tibble: 1 × 4  
## species\_common\_name species\_scientific\_name species\_abbreviat… total\_count  
## <chr> <chr> <chr> <dbl>  
## 1 Wedge-tailed shearwater Puffinus pacificus PUFPAC 855

# Answer: A Wedge-tailed shearwater, count 855

### Task 4 How many different types of birds were only ever seen in groups of 1?

# ok so theres got to be a better way but i ended up adding two columns on  
# with the individual sightings and the total count. If these were equal with no  
# NA values which were dropped then all must have been sightings of one bird  
  
seabirds\_counting <- seabirds %>%   
 drop\_na(count) %>%   
 group\_by(species\_common\_name) %>%   
 summarise(count\_of\_sightings = n(),   
 total\_count = sum(count, na.rm = TRUE))   
   
seabirds\_plus\_counts <- left\_join(seabirds,   
 seabirds\_counting,  
 "species\_common\_name" )  
  
task\_four\_data <- seabirds\_plus\_counts %>%   
 filter(count\_of\_sightings == total\_count) %>%   
 group\_by(species\_common\_name) %>%   
 summarise(n()) %>%   
 nrow()  
task\_four\_data

## [1] 86

### Task 5 How many penguins were seen

# Note here I've calculated count of penguins seen not the number of sightings  
  
# here we filter for any names containing penguin then we count the total for the count column  
seabirds %>%   
 filter(str\_detect(species\_common\_name,  
 "(?i).\*penguin+.\*")) %>%   
 summarise(sum(count, na.rm = T))

## # A tibble: 1 × 1  
## `sum(count, na.rm = T)`  
## <dbl>  
## 1 158

# 158 penguins were seen