Kelsey\_Data\_Cleaning

2023-03-01

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyr)  
library(readr)

catch\_original <- read.csv("https://knb.ecoinformatics.org/knb/d1/mn/v2/object/df35b.302.1")

“select” command from dplyr package selects *columns*; the filter command filters rows

## Clean Data

# Remove the All and notesRegCode columns from teh catch\_original dataframe  
catch\_data <- catch\_original %>%  
 select(Region, Year, Chinook, Sockeye, Coho, Pink, Chum)  
  
head(catch\_data)

## Region Year Chinook Sockeye Coho Pink Chum  
## 1 SSE 1886 0 5 0 0 0  
## 2 SSE 1887 0 155 0 0 0  
## 3 SSE 1888 0 224 16 0 0  
## 4 SSE 1889 0 182 11 92 0  
## 5 SSE 1890 0 251 42 0 0  
## 6 SSE 1891 0 274 24 0 0

glimpse(catch\_data)

## Rows: 1,708  
## Columns: 7  
## $ Region <chr> "SSE", "SSE", "SSE", "SSE", "SSE", "SSE", "SSE", "SSE", "SSE",…  
## $ Year <int> 1886, 1887, 1888, 1889, 1890, 1891, 1892, 1893, 1894, 1895, 18…  
## $ Chinook <chr> "0", "0", "0", "0", "0", "0", "0", "0", "0", "3", "4", "5", "9…  
## $ Sockeye <int> 5, 155, 224, 182, 251, 274, 207, 189, 253, 408, 989, 791, 708,…  
## $ Coho <int> 0, 0, 16, 11, 42, 24, 11, 1, 5, 8, 192, 161, 132, 139, 84, 107…  
## $ Pink <int> 0, 0, 0, 92, 0, 0, 8, 187, 529, 606, 996, 2218, 673, 1545, 204…  
## $ Chum <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 7, 0, 1, 2, 0, 0, 0, 102, 343…

#glimpse shows us that 'chinook' data are characters

#Mutate chinook class from character to numeric  
catch\_clean <- catch\_data %>%  
 mutate(Chinook = as.numeric(Chinook))

## Warning: There was 1 warning in `mutate()`.  
## ℹ In argument: `Chinook = as.numeric(Chinook)`.  
## Caused by warning:  
## ! NAs introduced by coercion

# search for the na value  
i <- which(is.na(catch\_clean$Chinook))  
i

## [1] 401

#when I call "i", it tells me the NA is on row 401  
#Turns out that the "I" should have been a "1"  
  
catch\_data[i, ]

## Region Year Chinook Sockeye Coho Pink Chum  
## 401 GSE 1955 I 66 0 0 1

catch\_clean <- catch\_data %>%  
 mutate(Chinook = if\_else(Chinook == "I", "1", Chinook)) %>%  
 mutate(Chinook = as.integer(Chinook))  
#for this purpose, "as.integer" and "as.numeric" do the same thing - but not if the numeric data are not integers  
  
glimpse(catch\_clean)

## Rows: 1,708  
## Columns: 7  
## $ Region <chr> "SSE", "SSE", "SSE", "SSE", "SSE", "SSE", "SSE", "SSE", "SSE",…  
## $ Year <int> 1886, 1887, 1888, 1889, 1890, 1891, 1892, 1893, 1894, 1895, 18…  
## $ Chinook <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 4, 5, 9, 12, 0, 3, 4, 0, 4, 9, 3…  
## $ Sockeye <int> 5, 155, 224, 182, 251, 274, 207, 189, 253, 408, 989, 791, 708,…  
## $ Coho <int> 0, 0, 16, 11, 42, 24, 11, 1, 5, 8, 192, 161, 132, 139, 84, 107…  
## $ Pink <int> 0, 0, 0, 92, 0, 0, 8, 187, 529, 606, 996, 2218, 673, 1545, 204…  
## $ Chum <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 7, 0, 1, 2, 0, 0, 0, 102, 343…

## Change shape fo data using pivot longer and pivot wider

# pivot longer practice  
catch\_long <- catch\_clean %>%  
 pivot\_longer(cols = -c(Region, Year), names\_to = "species", values\_to = "catch")  
catch\_long

## # A tibble: 8,540 × 4  
## Region Year species catch  
## <chr> <int> <chr> <int>  
## 1 SSE 1886 Chinook 0  
## 2 SSE 1886 Sockeye 5  
## 3 SSE 1886 Coho 0  
## 4 SSE 1886 Pink 0  
## 5 SSE 1886 Chum 0  
## 6 SSE 1887 Chinook 0  
## 7 SSE 1887 Sockeye 155  
## 8 SSE 1887 Coho 0  
## 9 SSE 1887 Pink 0  
## 10 SSE 1887 Chum 0  
## # … with 8,530 more rows

# pivot wider practice  
catch\_wide <- catch\_long %>%   
 pivot\_wider(names\_from = species, values\_from = catch)  
  
head(catch\_wide)

## # A tibble: 6 × 7  
## Region Year Chinook Sockeye Coho Pink Chum  
## <chr> <int> <int> <int> <int> <int> <int>  
## 1 SSE 1886 0 5 0 0 0  
## 2 SSE 1887 0 155 0 0 0  
## 3 SSE 1888 0 224 16 0 0  
## 4 SSE 1889 0 182 11 92 0  
## 5 SSE 1890 0 251 42 0 0  
## 6 SSE 1891 0 274 24 0 0

## Renaming columns

catch\_long <- catch\_long %>%   
 rename(catch\_thousands = catch)  
  
head(catch\_long)

## # A tibble: 6 × 4  
## Region Year species catch\_thousands  
## <chr> <int> <chr> <int>  
## 1 SSE 1886 Chinook 0  
## 2 SSE 1886 Sockeye 5  
## 3 SSE 1886 Coho 0  
## 4 SSE 1886 Pink 0  
## 5 SSE 1886 Chum 0  
## 6 SSE 1887 Chinook 0

## Add columns with mutate function

catch\_long <- catch\_long %>%   
 mutate(catch = catch\_thousands \* 1000)  
  
head(catch\_long)

## # A tibble: 6 × 5  
## Region Year species catch\_thousands catch  
## <chr> <int> <chr> <int> <dbl>  
## 1 SSE 1886 Chinook 0 0  
## 2 SSE 1886 Sockeye 5 5000  
## 3 SSE 1886 Coho 0 0  
## 4 SSE 1886 Pink 0 0  
## 5 SSE 1886 Chum 0 0  
## 6 SSE 1887 Chinook 0 0

#remove catch\_thousands  
catch\_long <- catch\_long %>%  
 mutate(catch = catch\_thousands \* 1000) %>%  
 select(-catch\_thousands)  
  
head(catch\_long)

## # A tibble: 6 × 4  
## Region Year species catch  
## <chr> <int> <chr> <dbl>  
## 1 SSE 1886 Chinook 0  
## 2 SSE 1886 Sockeye 5000  
## 3 SSE 1886 Coho 0  
## 4 SSE 1886 Pink 0  
## 5 SSE 1886 Chum 0  
## 6 SSE 1887 Chinook 0

## Practice using group\_by an summarize

mean\_region <- catch\_long %>%  
 group\_by((Region)) %>%  
 summarize(catch\_mean = mean(catch))  
  
head(mean\_region)

## # A tibble: 6 × 2  
## `(Region)` catch\_mean  
## <chr> <dbl>  
## 1 ALU 40384.  
## 2 BER 16373.  
## 3 BRB 2709796.  
## 4 CHG 315487.  
## 5 CKI 683571.  
## 6 COP 179223.

# summarize count by region (number of observations by region)  
n\_region <- catch\_long %>%   
 group\_by(Region) %>%  
 summarize(number\_of\_observations = n())  
  
head(n\_region)

## # A tibble: 6 × 2  
## Region number\_of\_observations  
## <chr> <int>  
## 1 ALU 435  
## 2 BER 510  
## 3 BRB 570  
## 4 CHG 550  
## 5 CKI 525  
## 6 COP 470

## Lesson on filtering

ALU\_catch <- catch\_long %>%  
 filter(Region == "ALU")  
# for two observations (filter(Region %n% c("ALU", "SSE"))  
# %n% is the "in" operator  
  
head(ALU\_catch)

## # A tibble: 6 × 4  
## Region Year species catch  
## <chr> <int> <chr> <dbl>  
## 1 ALU 1911 Chinook 0  
## 2 ALU 1911 Sockeye 9000  
## 3 ALU 1911 Coho 0  
## 4 ALU 1911 Pink 0  
## 5 ALU 1911 Chum 0  
## 6 ALU 1912 Chinook 0

## Lesson on arranging some rows

mean\_region <- catch\_long %>%   
 group\_by(Region) %>%   
 summarise(mean\_catch = mean(catch)) %>%   
 arrange(mean\_catch)  
  
head(mean\_region)

## # A tibble: 6 × 2  
## Region mean\_catch  
## <chr> <dbl>  
## 1 BER 16373.  
## 2 KTZ 18836.  
## 3 ALU 40384.  
## 4 NRS 51503.  
## 5 KSK 67642.  
## 6 YUK 68646.

# to put data in decending order, add "-"

# Practicing joins

region\_defs <- read.csv("https://knb.ecoinformatics.org/knb/d1/mn/v2/object/df35b.303.1") %>%   
 select(code, mgmtArea)  
  
head(region\_defs)

## code mgmtArea  
## 1 GSE Unallocated Southeast Alaska  
## 2 NSE Northern Southeast Alaska  
## 3 SSE Southern Southeast Alaska  
## 4 YAK Yakutat  
## 5 PWSmgmt Prince William Sound Management Area  
## 6 BER Bering River Subarea Copper River Subarea

## Join our data

catch\_joined <- left\_join(catch\_long, region\_defs, by = c("Region" = "code"))  
  
head(catch\_joined)

## # A tibble: 6 × 5  
## Region Year species catch mgmtArea   
## <chr> <int> <chr> <dbl> <chr>   
## 1 SSE 1886 Chinook 0 Southern Southeast Alaska  
## 2 SSE 1886 Sockeye 5000 Southern Southeast Alaska  
## 3 SSE 1886 Coho 0 Southern Southeast Alaska  
## 4 SSE 1886 Pink 0 Southern Southeast Alaska  
## 5 SSE 1886 Chum 0 Southern Southeast Alaska  
## 6 SSE 1887 Chinook 0 Southern Southeast Alaska

#Any Region "SSE" also gets "Southern Southeast Alaska"

region\_defs <- region\_defs %>%  
 rename(Region = code, Region\_Name = mgmtArea)  
  
catch\_joined <- left\_join(catch\_long, region\_defs, by = "Region")  
  
head(catch\_joined)

## # A tibble: 6 × 5  
## Region Year species catch Region\_Name   
## <chr> <int> <chr> <dbl> <chr>   
## 1 SSE 1886 Chinook 0 Southern Southeast Alaska  
## 2 SSE 1886 Sockeye 5000 Southern Southeast Alaska  
## 3 SSE 1886 Coho 0 Southern Southeast Alaska  
## 4 SSE 1886 Pink 0 Southern Southeast Alaska  
## 5 SSE 1886 Chum 0 Southern Southeast Alaska  
## 6 SSE 1887 Chinook 0 Southern Southeast Alaska

## Practice with separate and unite

sites\_df <- data.frame(site = c("HAW-101",  
 "HAW-103",  
 "OAH-320",  
 "OAH-219",  
 "MAI-039"))  
  
sites\_df %>%   
 separate(site, c("island", "site\_number"), "-")

## island site\_number  
## 1 HAW 101  
## 2 HAW 103  
## 3 OAH 320  
## 4 OAH 219  
## 5 MAI 039

dates\_df <- data.frame(year = c("1930",  
 "1930",  
 "1930"),  
 month = c("12",  
 "12",  
 "12"),  
 day = c("14",  
 "15",  
 "16"))  
  
dates\_df %>%   
 unite(date, year, month, day, sep = "-")

## date  
## 1 1930-12-14  
## 2 1930-12-15  
## 3 1930-12-16