Data Cleaning and Manipulation

Anastasija Mensikova

```
library(dplyr)
library(tidyr)

# To call stats filter -- stats::filter()
```

Reformat the catch data

- Remove "all" column
- Remove "notes" column
- Create a "species" column
  - Move from wide to long
- General QA

Mike Byerly. Alaska commercial salmon catches by management region (1886- 1997). Gulf of Alaska Data Portal. df35b.304.2.

```
catch_original = read.csv("https://knb.ecoinformatics.org/knb/d1/mn/v2/object/df35b.302.1", stringsAsFa

head(catch_original)
```

```
##   Region Year Chinook Sockeye Coho Pink Chum All notesRegCode
## 1    SSE 1886       0       5    0    0    0   5
## 2    SSE 1887       0     155    0    0    0 155
## 3    SSE 1888       0     224   16    0    0 240
## 4    SSE 1889       0     182   11   92    0 285
## 5    SSE 1890       0     251   42    0    0 292
## 6    SSE 1891       0     274   24    0    0 298
```

Remove the "all" and "notesRegCode" columns using "select"

```
#Cmd + shift + m: pipe operator shortcut ( %>% )

# catch_long = catch_original %>% select(-All, -notesRegCode) %>% ... --> this will be the same

catch_long = catch_original %>% select(Region, Year, Chinook, Sockeye, Coho, Pink, Chum) %>%
                                gather(key = "species", value = "catch", -Region, -Year)

head(catch_long)
```

```
##   Region Year species catch
## 1    SSE 1886 Chinook     0
## 2    SSE 1887 Chinook     0
## 3    SSE 1888 Chinook     0
## 4    SSE 1889 Chinook     0
## 5    SSE 1890 Chinook     0
## 6    SSE 1891 Chinook     0
```

```
catch_wide = catch_long %>% spread(key = species, value = catch)

head(catch_wide)
```

```
##   Region Year Chinook Chum Coho Pink Sockeye
## 1    ALU 1911       0    0    0    0       9
## 2    ALU 1912       0    0    0    0       0
## 3    ALU 1913       0    0    0    0       0
## 4    ALU 1914       0    0    0    0       0
## 5    ALU 1915       0    0    0    0       0
## 6    ALU 1916       0    0    1  180      76
```

Clean up our data

- Rename catch to catch_thousands
- Change "catch" column to numeric
- Create a new "catch" column in units num. of fish

```
catch_clean = catch_long %>% rename(catch_thousands = catch) %>%
                             mutate(catch_thousands = ifelse(catch_thousands == "I", 1, catch_thousands)
                             mutate(catch_thousands = as.numeric(catch_thousands)) %>%
                             mutate(catch = catch_thousands * 1000) %>%
                             select(-catch_thousands)

head(catch_clean)
```

```
##   Region Year species catch
## 1    SSE 1886 Chinook     0
## 2    SSE 1887 Chinook     0
## 3    SSE 1888 Chinook     0
## 4    SSE 1889 Chinook     0
## 5    SSE 1890 Chinook     0
## 6    SSE 1891 Chinook     0
```

Split - Apply - Combine

- Calculate mean catch by species

```
species_mean = catch_clean %>% group_by(species, Region) %>% summarise(catch_mean = mean(catch), num_obs

head(species_mean)
```

```
## # A tibble: 6 x 4
## # Groups:   species [1]
##   species Region catch_mean num_obs
##   <chr>   <chr>       <dbl>   <int>
## 1 Chinook ALU          23.0      87
## 2 Chinook BER          19.6     102
## 3 Chinook BRB       76211.      114
## 4 Chinook CHG        1536.      110
## 5 Chinook CKI       43876.      105
## 6 Chinook COP       19798.       94
```

Join the region definitions

```
regions_original = read.csv("https://knb.ecoinformatics.org/knb/d1/mn/v2/object/df35b.303.1")
```

Miscellaneous functions