Data Tidying

Shayne O'Brien

January 16, 2019

```
library(tidyr)
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
```

To call a funtion from a specific package 'package_name::function_name(…) This is in the case of overlap in objects per the above error message

To Hide, use `{r, warning= FALSE, message = False}`

Data Cleaning

Pipe Operator (%>%)

The Pipe Operator effieciently chains operations together.

Use: [Ctrl+Shift+M]

Practice

```
catch_df<- read.csv(url("https://knb.ecoinformatics.org/knb/d1/mn/v2/object/df35b.302.1",
                        method = "libcurl"),
                        stringsAsFactors = FALSE)
```

The above code allows reading in data from a url. `read.csv(file =)` sometimes doesnt work on windows, the above code fixes the error.

`libcurl` forces the default library to make a connection with an https:// URL. Dependent on operating system.

```
head(catch_df)
```

```
##   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
```

```
catch_long<- catch_df %>%
  select(Region, Year, Chinook, Sockeye, Coho, Pink, Chum) %>%
  gather(key = "Species", value = "catch", Chinook, Sockeye, Coho, Pink, Chum)


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
```

8erroneus value due to OCR issue - Change “I” to one *create catch column in correct units

```
catch_cleaned<-catch_long %>%
  rename(catch_thousands = catch)  %>%
  mutate(catch_thousands = ifelse(catch_thousands == "I", 1, catch_thousands)) %>%
  mutate(catch_thousands = as.integer(catch_thousands)) %>%
  mutate(catch = catch_thousands * as.integer(1000))

tail(catch_cleaned)
```

```
##      Region Year Species catch_thousands  catch
## 8535    NOP 1992    Chum             342 342000
## 8536    NOP 1993    Chum             135 135000
## 8537    NOP 1994    Chum              84  84000
## 8538    NOP 1995    Chum              99  99000
## 8539    NOP 1996    Chum              68  68000
## 8540    NOP 1997    Chum              97  97000
```

Split-Apply-Combine

Calulculate total catch by region

```
catch_total <- catch_cleaned %>%
  group_by(Region) %>%
  summarize(catch_region = mean(catch))
            #n_obs = n())

catch_total
```

```
## # A tibble: 18 x 2
##    Region catch_region
##    <chr>         <dbl>
##  1 ALU          40384.
##  2 BER          16373.
##  3 BRB        2709796.
##  4 CHG         315487.
##  5 CKI         683571.
##  6 COP         179223.
##  7 GSE         133841.
##  8 KOD        1528350
##  9 KSK          67642.
## 10 KTZ          18836.
## 11 NOP         229493.
## 12 NRS          51503.
## 13 NSE        1825021.
## 14 PWS        1419237.
## 15 SOP        1110942.
## 16 SSE        3184661.
## 17 YAK          91923.
## 18 YUK          68646.
```

Joins