Exploring Global Fisheries Collapse Using dplyr

Claire Parkinson

October 30, 2017

Introduction

In class, we have learned many of the most popular functions in the R package "dplyr" in order to perform basic data manipulations. The dplyr package is extremely useful when analyzing data in R, and is frequently used in the data science world. The purpose of this assignment is to write a report, in the form of a blog post, about one or more topics covered so far in the course, in greater depth, displaying mastery of the material. For my post, I have chosen to explore data and table manipulation, with a special focus on the dyplr package, but we will also use other tidyverse packages like readxl and ggplot. To do this, I have used data on global fisheries from the RAM Legacy Stock Assessment Database (cited in References) in accordance with their Fair Use Policy. We will explore how the total catch of fish and the percentage of collapsed fisheries has changed over time, as well as practice different data manipulation techniques using dplyr functions.

Summary of dplyr Functions

We will use almost all of these functions in our data analysis. Many of these functions we have practiced in class, but there are also some here that we haven't gone over, but which I learned while doing this assignment. I will explain with comments in my code how I used these functions in my data analysis.

- *slice(table, 1:4)*
  - Selects first four rows of table
  - Can put any number or range of rows
- *select(table, column1, column2)*
  - Selects columns 1 and 2 from table
  - No need to put quotes around column names
- *filter(table, column (condition) value)*
  - Filters rows by condition
  - Condition can include nequalitites like ">","<","==", etc.
  - Condition "%in%" filters by a specific vector of values
- *mutate(table, column_name = values)*
  - Adds column of values to table
- *arrange(table, column1)*
  - Sorts rows by column 1 in increasing order
- *arrange(table, desc(column1))*
  - Sorts rows by column 1 in decreasing order
- *group_by(table, column1)*
  - Groups data in table by column1 values
- *summarise(table, avgs = mean(column1))*
  - Gives the average of values in column 1
- *summarise(group_by(table, column1), avgs_by_column2 = mean(column2), na.rm = TRUE)*
  - Returns a table with averages of column 2 for each value in column 1
  - Can use any function besides mean
  - Adding na.rm = TRUE drops missing values before computation
- *rename(table, new_name = old_name)*
  - Renames variable old_name from table with new_name
- *distinct(table, variable)*
  - Returns only unique values of specified variable from table
- *unique(character)*
  - Returns only unique values in vector of characters
- *pull(table, variable)*
  - Extracts variable values from table
- *full_join(table1, table2, by = NULL)*
  - Joins table1 and table2 by the character vector specified
  - Returns all rows and columns in both tables, returns NA for nonmatching values
- *left_join(table1, table2, by = NULL)*
  - Joins table1 and table2 by the character vector specified
  - Returns all rows from table1 and all columns from both tables
  - Rows in table1 with no match in table2 will have NA values
- *right_join(table1, table2, by = NULL)*
  - Joins table1 and table2 by the character vector specified
  - Returns all rows from table2 and all columns from both tables
  - Rows in table2 with no match in table2 will have NA values

Examples Using Data

Data Preparation

```
# load necessary packages
library("readxl")
library("ggplot2")
```

```
## Warning: package 'ggplot2' was built under R version 3.4.2
```

```
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
```

```
# download the data and prepare it for analysis

download.file("https://depts.washington.edu/ramlegac/wordpress/databaseVersions/RLSADB_v3.0_(assessment_data_only)_excel.zip",
              "ramlegacy.zip") # download the data
path <- unzip("ramlegacy.zip")  # unzip the .xls files
sheets <- readxl::excel_sheets(path) # use the readxl package to identify sheet names
ram <- lapply(sheets, readxl::read_excel, path = path)  # read the data from all 3 sheets into a list
names(ram) <- sheets # give the list of datatables their assigned sheet names

sheets <- readxl::excel_sheets("RLSADB v3.0 (assessment data only).xlsx") # download the other data set

ram_data <- lapply(sheets, read_excel, path = "RLSADB v3.0 (assessment data only).xlsx") # read the data from all 3 sheets from this data set into a list
names(ram_data) <- sheets # give the list of this set of datatables their assigned sheet names

attach(ram_data) # attach names to database
```

Section 1: Basic Data Manipulation

Slicing Data

```
# Examine data table on types of fish

slice(stock, 1:10) # use slice() to look at first 10 rows
```

```
## # A tibble: 10 x 9
##          stockid    tsn                  scientificname      commonname
##            <chr>  <dbl>                           <chr>           <chr>
##  1 ACADREDGOMGB 166774             Sebastes fasciatus Acadian redfish
##  2       AFLONCH 166156                Beryx splendens       Alfonsino
##  3        ALBAIO 172419               Thunnus alalunga   albacore tuna
##  4       ALBAMED 172419               Thunnus alalunga   albacore tuna
##  5      ALBANATL 172419               Thunnus alalunga   Albacore tuna
##  6      ALBANPAC 172419               Thunnus alalunga   Albacore tuna
##  7      ALBASATL 172419               Thunnus alalunga   albacore tuna
##  8      ALBASPAC 172419               Thunnus alalunga   Albacore tuna
##  9  ALPLAICBSAI 172901 Pleuronectes quadrituberculatus   Alaska plaice
## 10       AMPL23K 172877    Hippoglossoides platessoides American Plaice
## # ... with 5 more variables: areaid <chr>, stocklong <chr>, region <chr>,
## #   inmyersdb <dbl>, myersstockid <chr>
```

Filtering and Extracting Data

```
# Only look at data for Atlantic Cod

stockids <- stock %>%
  filter(commonname == "Atlantic cod") %>%  # filter by Atlantic Cod
  pull(stockid) # Extract stockids for Atlantic Cod

stockids
```

```
##  [1] "COD2J3KL"  "COD3M"     "COD3NO"    "COD3Pn4RS" "COD3Ps"
##  [6] "COD4TVn"   "COD4VsW"   "COD4X"     "CODBA2224" "CODBA2532"
## [11] "CODFAPL"   "CODGB"     "CODGOM"    "CODICE"    "CODIS"
## [16] "CODKAT"    "CODNEAR"   "CODNS"     "CODVIa"
```

Section 2: Total Catch Over Time

Grouping and Summarizing Data

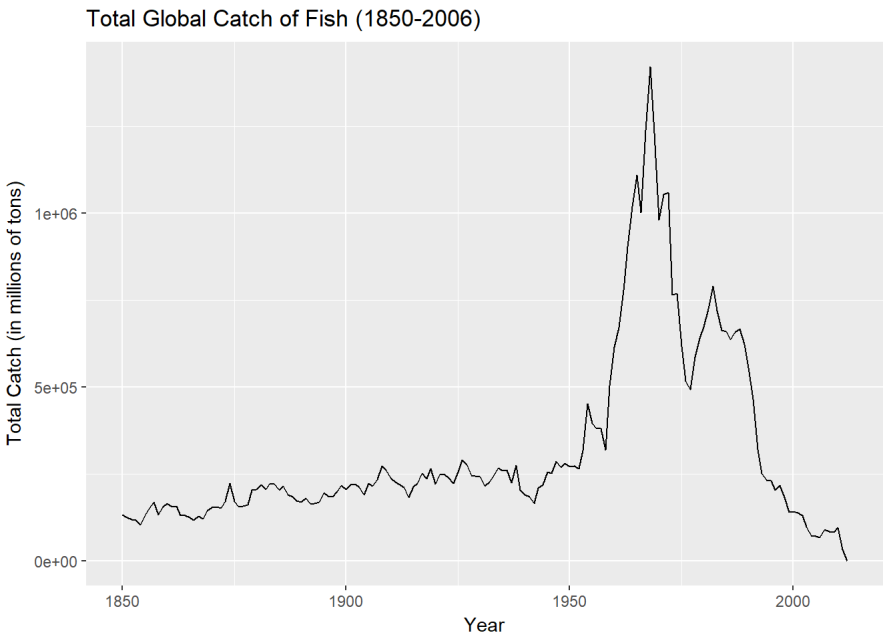
```
# Aggregating total catch of Atlantic Cod over time

total_catch <- timeseries_values_views %>% # use pipe operator to perform many operations on same table
  filter(stockid %in% stockids) %>%   # filter by fish with Atlantic Cod stockids
  group_by(year) %>%                  # group data by year
  summarize(catch = sum(TC, na.rm=TRUE))# summarize the total catch for each year
total_catch
```

```
## # A tibble: 163 x 2
##     year  catch
##    <dbl>  <dbl>
##  1  1850 133100
##  2  1851 125400
##  3  1852 120000
##  4  1853 116600
##  5  1854 103900
##  6  1855 131500
##  7  1856 150800
##  8  1857 169300
##  9  1858 133800
## 10  1859 153900
## # ... with 153 more rows
```

Graphing

```
# Graphing total catch of fish over time
total_catch %>%
  ggplot(aes(year, catch)) + ggtitle("Total Global Catch of Fish (1850-2006)") + geom_line() + ylab("Total Catch (
in millions of tons)") + xlab("Year")
```



This graph shows us that, after 100 years of stability, total global catch of fish exploded in the late 1950's, quickly peaking and falling very steeply until 1975. There was then a brief increase for several years before total catch fell sharply again, approaching total depletion in the early 2000's.

Renaming Variables

```
# rename variables with same names in different tables
units <- timeseries_units_views %>%
  select(stockid, SSB, TC) %>%
  rename(SSB_units = SSB) %>%
  rename(catch_landings_units = TC)
```

Joining Data Tables

```
# joining all data tables

# joining tables for all fish
fish <- timeseries_values_views %>%
  left_join(stock) %>%
  left_join(area) %>%
  left_join(units) %>%
  select(stockid, country, SSB_units, catch_landings_units, scientificname, commonname, year, SSB, TC, areaname)
```

```
## Joining, by = c("stockid", "stocklong")
```

```
## Joining, by = "areaid"
```

```
## Joining, by = "stockid"
```

```
slice(fish, 1:10) # only show first 10 rows of table
```

```
## # A tibble: 10 x 10
##          stockid country SSB_units catch_landings_units     scientificname
##            <chr>   <chr>     <chr>                <chr>              <chr>
##  1 ACADREDGOMGB     USA        MT                   MT Sebastes fasciatus
##  2 ACADREDGOMGB     USA        MT                   MT Sebastes fasciatus
##  3 ACADREDGOMGB     USA        MT                   MT Sebastes fasciatus
##  4 ACADREDGOMGB     USA        MT                   MT Sebastes fasciatus
##  5 ACADREDGOMGB     USA        MT                   MT Sebastes fasciatus
##  6 ACADREDGOMGB     USA        MT                   MT Sebastes fasciatus
##  7 ACADREDGOMGB     USA        MT                   MT Sebastes fasciatus
##  8 ACADREDGOMGB     USA        MT                   MT Sebastes fasciatus
##  9 ACADREDGOMGB     USA        MT                   MT Sebastes fasciatus
## 10 ACADREDGOMGB     USA        MT                   MT Sebastes fasciatus
## # ... with 5 more variables: commonname <chr>, year <dbl>, SSB <dbl>,
## #   TC <dbl>, areaname <chr>
```

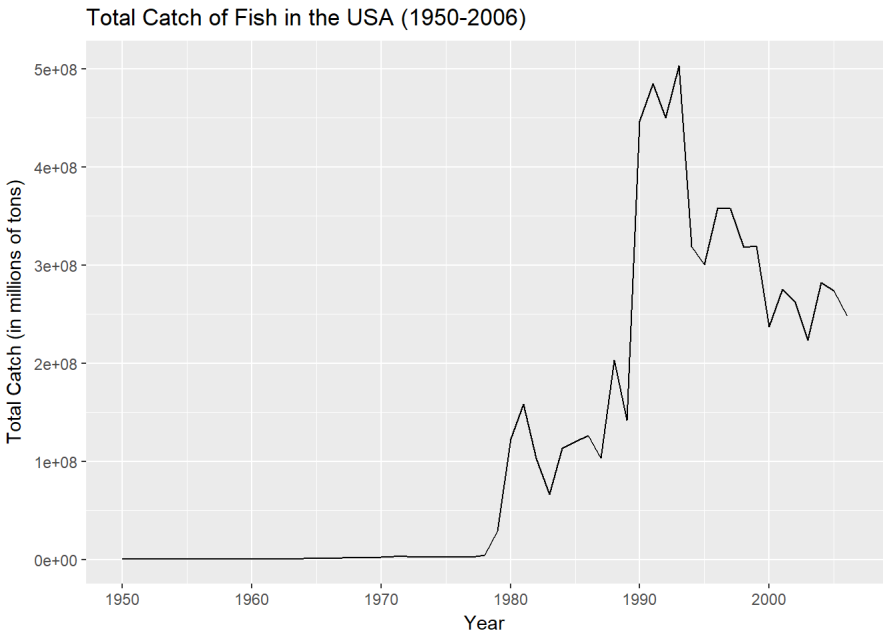
Subsetting Data

```
# subset data by country to look only at fish in the USA
usa_fish <- filter(fish, country == 'USA') # filter table by fish in the USA

usa_fish_ids <- unique(usa_fish[[1]]) # Extract all unique stockids for USA fish

# note: did not display usa_fish_ids since there are 159 of them
```

```
# plotting total catch over time of fish in the USA
usa_TC <- usa_fish %>%
  filter(year>=1950, year<=2006) %>%
  group_by(year) %>%
  summarise(catch = sum(TC, na.rm = TRUE)) %>%
  ggplot(aes(year, catch)) + geom_line() + ggtitle("Total Catch of Fish in the USA (1950-2006)") + ylab("Total Cat
ch (in millions of tons)") + xlab("Year")
usa_TC
```



Observing fishery data only in the USA, this graph shows us that total catch increased sharply in the 1990's and then fell by about half by 1975, then decreasing moderately and oscillating up and down in the 2000's.

Section 3: Global Fisheries Collapse Over Time

Data Preparation

```
# only include fish species with data for the full range of years
full_fish <- fish %>%
  filter(between(year, 1950, 2006)) %>% # filter by time frame
  group_by(stockid) %>% # aggregate data for each species of fish
  filter(TC != 'NA') %>% # drop species with NA values for total catch
  summarise(n_year = n()) %>% filter(n_year == 57) # make sure all species have 57 years of data

stockids_full_range <-full_fish$stockid #extract stockids with data for full range
```

Calculating Collapse Values

```
collapses <- fish %>%
  filter(stockid %in% stockids_full_range, year>=1950, year<=2006) %>% # filter by time frame, using stockids with full range of years
  select(stockid, TC, year) %>% # select only needed variables
  group_by(stockid) %>% # aggregate data for each species of fish (stockid)
  mutate(peak = max(TC, na.rm = TRUE), # calculate the peak population for each species
         min_collapse = 0.1*peak, # let the minimum collapse value be 10% of that species' peak
         collapsed = ifelse(TC < min_collapse, TRUE, FALSE), # return TRUE if species population is below the minimum collapse value (or else FALSE)
         cumulative = cumsum(collapsed)) # count the number of years the species has been collapsed for at that point in time
collapses
```

```
## # A tibble: 3,249 x 7
## # Groups:   stockid [57]
##          stockid    TC  year  peak min_collapse collapsed cumulative
##            <chr> <dbl> <dbl> <dbl>        <dbl>     <lgl>      <int>
##  1 ACADREDGOMGB 34307  1950 34307       3430.7     FALSE          0
##  2 ACADREDGOMGB 30077  1951 34307       3430.7     FALSE          0
##  3 ACADREDGOMGB 21377  1952 34307       3430.7     FALSE          0
##  4 ACADREDGOMGB 16791  1953 34307       3430.7     FALSE          0
##  5 ACADREDGOMGB 12988  1954 34307       3430.7     FALSE          0
##  6 ACADREDGOMGB 13914  1955 34307       3430.7     FALSE          0
##  7 ACADREDGOMGB 14388  1956 34307       3430.7     FALSE          0
##  8 ACADREDGOMGB 18490  1957 34307       3430.7     FALSE          0
##  9 ACADREDGOMGB 16047  1958 34307       3430.7     FALSE          0
## 10 ACADREDGOMGB 15521  1959 34307       3430.7     FALSE          0
## # ... with 3,239 more rows
```

Calculating Percentage of Fisheries Collapsed

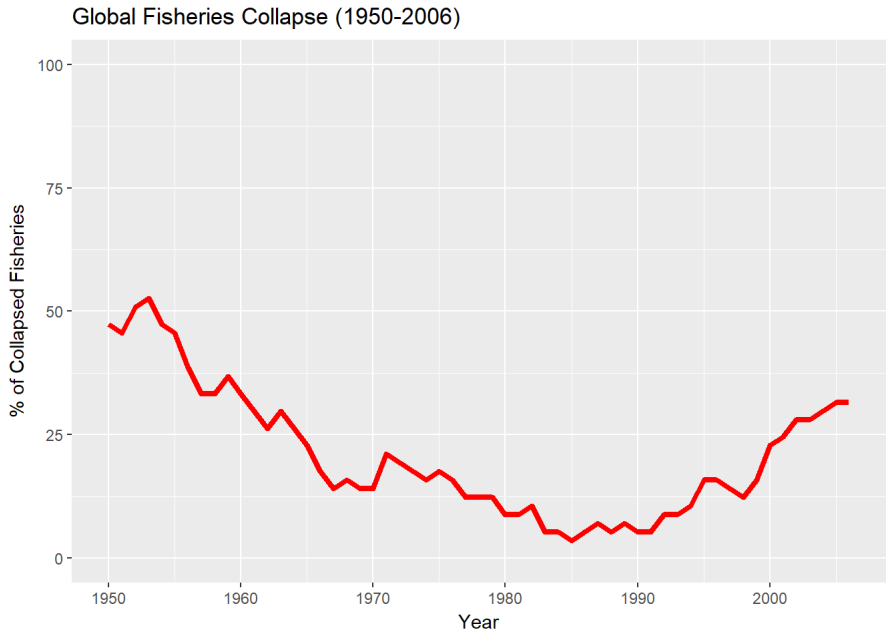
```
# table of percent of collapsed fisheries per year (1950-2006)
perc_collapsed <- summarise(
    group_by(collapses, year),
    percent_collapsed = sum(collapsed)/length(collapsed)*100)
perc_collapsed
```

```
## # A tibble: 57 x 2
##     year percent_collapsed
##    <dbl>             <dbl>
##  1  1950          47.36842
##  2  1951          45.61404
##  3  1952          50.87719
##  4  1953          52.63158
##  5  1954          47.36842
##  6  1955          45.61404
##  7  1956          38.59649
##  8  1957          33.33333
##  9  1958          33.33333
## 10  1959          36.84211
## # ... with 47 more rows
```

Graphing

```
# graph percentage of fisheries collapsed over time

perc_collapsed %>%
ggplot(aes(year, percent_collapsed)) + ggtitle("Global Fisheries Collapse (1950-2006)") + xlab("Year") + ylab("% of Collapsed Fisheries") + geom_line(lwd = 1.5, col = 'red') + ylim(c(0, 100))
```



This graph shows us that the percentage of collapsed fisheries steadily decreased from 1950 to the 1990's, but has started to quickly increase again since the 1990's and into the 2000's.

Discussion

On Data Analysis Methods

Using many of the functions from the dplyr package that we summarized in the Background section of this post, we were able to successfully manipulate and analyze our data on global fisheries. We saw through examples how we can make many complicated operations on data tables using dplyr functions like select(), filter(), group_by(), summarise(), left_join(), and mutate() and the pipe operator (%>%). We learned how to use some new dplyr functions that we have not covered or used much in class such as pull(), unique(), and the %in% condition in filter. We also learned how to and practiced using the ifelse() and cumsum() functions and plotted our data with ggplot. In our data preparation methods, we learned how to download, unzip, read in, and assign names to data formatted as Excel sheets using functions from the readxl package.

On Results

From our analysis of global fisheries collapse, we can see that the 1950's were a time of extreme overfishing that resulted in over 50% of global fisheries collapsing. Although, total catch fell significantly from this peak and many of the collapsed fisheries were able to recover, recent trends since the 1990's are worrisome. While aggregated across the entire world's fishing industry, total catch has not significantly increased, total catch in the USA has been moderately increasing, and the percentage of collapsed fisheries has started increasing rapidly after 50 years of declining with the most recent data point in 2006 indicating that 30% of fisheries collapsed.

Conclusion

There are many useful functions in the dplyr package for performing a variety of data analyses, as we showed using global fishery data as an example. These functions are fairly easy to use and learn and when coupled with Base R functions and other tidyverse packages like ggplot, become very powerful tools for understanding, visualizing, and analyzing data.

References

Data:

Ricard, D., Minto, C., Jensen, O.P. and Baum, J.K. (2013) Evaluating the knowledge base and status of commercially exploited marine species with the RAM Legacy Stock Assessment Database. Fish and Fisheries 13 (4) 380-398. DOI: 10.1111/j.1467-2979.2011.00435.x

Supplementary Sources:

Dean, Cornelia. "Study Sees 'Global Collapse' of Fish Species." The New York Times, The New York Times, 2 Nov. 2006. (http://nytimes.com/2006/11/03/science/03fish.html?_r=0).

Grolemand, Garrett, and Hadley Wickham. R for Data Science. Online textbook. (http://r4ds.had.co.nz/#).

Impacts of Biodiversity Loss on Ocean Ecosystem Services. BY BORIS WORM, EDWARD B. BARBIER, NICOLA BEAUMONT, J. EMMETT DUFFY, CARL FOLKE, BENJAMIN S. HALPERN, JEREMY B. C. JACKSON, HEIKE K. LOTZE, FIORENZA MICHELI, STEPHEN R. PALUMBI, ENRIC SALA, KIMBERLEY A. SELKOE, JOHN J. STACHOWICZ, REG WATSON SCIENCE03 NOV 2006 : 787-790. (http://science.sciencemag.org/content/314/5800/787).

"Manipulating data with dpylr." Sparklyr, RStudio, 2016. (http://spark.rstudio.com/dplyr.html).

Millennium Ecosystem Assessment. 2005. Summary for Decision Makers. In Ecosystems and Human Well-being: Synthesis, 1-24. Washington, D.C.: Island Press.(https://groups.nceas.ucsb.edu/sustainability-science/2010%20weekly-sessions/session-5-2013-10.11.2010-the-environmental-services-that-flow-from-natural-capital/supplemental-readings-from-the-reader/MEA%20synthesis%202005.pdf/at_download/file).

Wickham, Hadley. "Dpylr: A Grammar of Data Manipulation." RDocumentation, RDocumentation, 24 June 2016, 15:37:11. (http://rdocumentation.org/packages/dplyr/versions/0.5.0.).