Data Management and Transformation

Juan Carlos Villaseñor-Derbez (JC)

Exercise 1: Reading and transforming data

Your boss asked you what sounds like a simple query:

> How much money have tuna purse seiners made since 2000 when fishing for bigeye tuna (*Thunnus obesus*) in the Eastern Pacific Ocean?

Let's make some assumptions that will help us answer this question:

1. We will interpret "making money" as revenue, not profits
2. The market price of tuna since 2000 has remained relatively stable, at around US$2/Kg (See Sibert et al. (2012))
3. We will focus on tuna production in the Eastern Pacific Ocean as reported by the IATTC

~~Part A) Obtaining data from the wild~~ (done on Tuesday)

How to find the data:

1. Go to iattc.org
2. In the top menu, **hover** over `DATA`
3. Click on "Public domain"
4. You will be taken to a page titled "Public domain data for download"
5. We will use "EPO total estimated catch by year, flag, gear, species"
6. Click on `CatchByFlagGear.zip` to the right of the table to prompt a download
7. Save the zip file to inside your EVR628 project at: `data/raw/`[1]
8. Using your finder / explorer, navigate to `EVR628/data/raw/` and unzip the `CatchByFlagGear.zip` file[2]
9. You will get a new folder called `CatchByFlagGear`
10. Read the PDF file enclosed, which contains the documentation

[1] If your web browser didn't allow you to specify the download folder, your file is likely in the "Downloads" folder. Navigate there and copy it to the `data/raw/` folder.

[2] Windows users: You might have to click a button called "Extract" in the top of your explorer window.

Part B) Reading data

0. **Put your post-it up**
1. Start a new script called `tuna_analysis` and save it to your `scripts/03_analysis` folder
2. Add a comment header and outline, and load the `tidyverse` and `EVR628tools` packages at the top of your script
3. Use the `read_csv()` function to load the new data and assign it to an object called `tuna_data`
4. What are the existing column names?[3]
5. **Remove your post-it when you are done**

```
# My code will be here
```

Part C) Renaming columns with `clean_names()` and `rename()`

0. **Put your post-it up**
1. In your console, install the `janitor` package using `install.packages("janitor")`
2. Load `janitor` at the top of your script, and then read the documentation for the `clean_names()` function
3. Modify your code for the `tuna_data` object so that you pipe into `clean_names()` after reading the data
4. What are the new column names?
5. Extend the pipeline above so that we can use the `rename()` function. Rename the columns so that we only retain the English portion of the name. Let's also rename `t` as `catch`
6. **Remove your post-it when you are done**

Congratulations, you now have a tidy data set with which we can work! The next steps are to keep the data we care about, calculate revenues, and then calculate summaries. Let's do that.

Part D) Filtering rows with `filter()`

0. **Put your post-it up**
1. What are the unique species represented in the data?
2. What are the fishing gears represented in the data?
3. Does the documentation say what these codes are?[4]
4. What is the species code for bigeye tuna?
5. What is the gear code for purse seine?
6. Create a new object called `ps_tuna_data` that takes the `tuna_data` and filters it to retain data for:
   a. bigeye tuna
   b. caught by tuna purse seiners

[3] Hint: use the `colnames()` function
[4] Hint: There is a cryptic link to the reference codes

   c. since 2000
7. **Remove your post-it when you are done**

Part E) Creating new columns with `mutate()`

0. **Put your post-it up**
1. Modify the `ps_tuna_data` pipeline to create a new column called `revenue` that calculates the revenue generated by selling the catch[5]
2. Make sure you calculate revenues in Millions of USD
3. **Remove your post-it when you are done**

```
# My code will be here
```

Part F) Calculating group summaries with `group_by()` and `summarize()`

0. **Put your post-it up**
1. The data right now report catch at the year-by-flag level. Modify the `ps_tuna_data` pipeline so that we have total catch and revenue by year.[6]
2. **Remove your post-it when you are done**

```
# My code will be here
```

Part G) Visualize the data and answer the question

Remember, the question was:

> How much money have tuna purse seiners made since 2000 when fishing for bigeye tuna (*Thunnus obesus*)?

The question is ambiguous because one could answer "They have made X M USD since 2000" or "Every year since 2000, they have made Y M USD per year." So let's get both:

0. **Put your post-it up**
1. What is the total revenue?[7]
2. What is the average annual revenue?[8]
3. Build a time-series showing revenues by year. Make sure to correctly label the axis and include a title and caption describing the figure and the data source, respectively.
4. **Remove your post-it when you are done.**

```
# My code will be here
```

[5]Hint: If I catch 10 kilos and the price per kilo is US$2, then I make US$20 because $10 * 2 = 20$

[6]Hint: You will need to use the `group_by`, `summarize()`, and `sum()` functions.

[7]Hint: Use `$` and `sum()`

[8]Hint: Use `$` and `mean()`

Extra exercises for you to practice

1. Make a figure showing total catch by species during 2023
2. Which country caught the most tuna during 2020?
3. For each species, identify the year in which catch was at it's maximum
4. How many species have been caught by Mexican-flagged vessels since 2000?

Sibert, John, Inna Senina, Patrick Lehodey, and John Hampton. 2012. "Shifting from Marine Reserves to Maritime Zoning for Conservation of Pacific Bigeye Tuna (Thunnus Obesus)." *Proc. Natl. Acad. Sci. U. S. A.* 109 (October): 18221–25.