SM-2302 Labs (R3)

1. The `palmerpenguin` package contains measurement data on various penguin species on islands near Palmer Station in Antarctica. The code below shows the number of each species measured on each of the three islands (missing island-penguin pairs implies that species does not occur on that island).

```
palmerpenguins::penguins %>%
  count(island, species)
```

```
## # A tibble: 5 x 3
##   island    species       n
##   <fct>     <fct>     <int>
## 1 Biscoe    Adelie       44
## 2 Biscoe    Gentoo      124
## 3 Dream     Adelie       56
## 4 Dream     Chinstrap    68
## 5 Torgersen Adelie       52
```

   Starting from these data construct a contingency table of counts for island (rows) by species (columns) using the pivot functions we've just discussed.

2. The Gapminder foundation is an independent educational non-profit organisation that promotes sustainable global development by increased use and understanding of statistics and other socio-economic information. One of the co-founders, Hans Rosling, was a great promoter of the use of data and visualisation to educate people about current world issues. His BBC Four video remains one of my favourite data visualisation videos ever, along with his other TED talks.

   The file `gapminder_fertility.csv` contains an excerpt from the larger Gapminder data set. The CSV file focuses on fertility rates (average number of children per woman) from 1960 to 2015 for (almost) all countries.

   (a) Load the data set in R. The data set is currently in *wide* format. Use the `pivot_*` functions to convert the data set to *long* format.

   (b) Summarise the data set to obtain the average fertility rate for each continent for years before and after the year 1990.

3. This question pertains to the `nycflights13::flights` data set.

   (a) How many flights to Los Angeles (LAX) did each of the legacy carriers (AA, UA, DL or US) have in May from JFK, and what was their average duration?

   (b) What was the shortest flight out of each airport in terms of distance? In terms of duration?

   (c) Which plane (check the tail number) flew out of each New York airport the most?

   (d) Which date should you fly on if you want to have the lowest possible average departure delay? What about arrival delay?

4. `repurrrsive` is an R package that contains a number of interesting example data sets that are stored in a hierarchical format. Many come from web-based APIs which provide results as JSON. We'll be looking at the following data sets pertaining to the Star Wars Universe:

   - `sw_people`
   - `sw_starships`

- `sw_films`
- `sw_planet`

(a) Before starting, go ahead and read the help file on these data sets first. Specifically, what is the object type? What are their dimensions/length? *Hint: Use the `View()` function on these objects.*

(b) Clearly, the `sw_*` objects are not suitable for data manipulation. So, we'll convert them to tibbles. Run the following commands in R:

```
library(repurrrsive)

# Star Wars people dataset
people <- tibble(people = sw_people) %>%
  unnest_wider(people) %>%
  unnest_longer(starships)

# Star Wars ship dataset
ships <- tibble(ships = sw_starships) %>%
  unnest_wider(ships)
```

Inspect the newly created tibbles. What do the `unnest_wider()` and `unnest_longer()` functions do?

(c) Amend the above the code, by adding more pipes, so that the following are achieved:

- `people`: Keep only the `name` of the people and the column `starships`.
- `ships`: Keep only the name of the ship (rename this column to be `ship`) and the `url` column.

Using the appropriate joining functions (`left_join()`, `inner_join()`, etc.–see `https://github.com/gadenbuie/tidyexplain`), **merge the `people` and `ships` tibbles together**. *Hint: Find the common column between the two tibbles, and use that as the `by` argument in the join functions.*

Finally, trim down and summarise the joined tibble so that what remains is a `20 x 2` tibble containing the name of the character, and all the ships that character has piloted. *Hint: Use the `group_by()` and `summarise()` functions, among other things.*

You should get some thing that looks like this:

```
## # A tibble: 20 x 2
##   name              ships
##   <chr>             <chr>
## 1 Anakin Skywalker  Trade Federation cruiser, Jedi Interceptor, Naboo~
## 2 Arvel Crynyd      A-wing
## 3 Biggs Darklighter X-wing
## 4 Boba Fett         Slave 1
## 5 Chewbacca         Millennium Falcon, Imperial shuttle
## # ... with 15 more rows
```

(Optional) More practice: Answer the following questions.

1. Which planet appeared in the most starwars film (according to the data in `sw_planet`)?
2. Which planet was the home world of the most characters in the Star Wars films?