Data Manipulation Assignment

Promise Idahosa

INTRODUCTION

Running Code

When you click the **Render** button a document will be generated that includes both content and the output of embedded code. You can embed code like this:

```r
1 + 1
```

```
[1] 2
```

You can add options to executable code like this

```
[1] 4
```

```r
# Load necessary libraries
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
```

```r
library(readr)

# Import the data
asfis_data <- read_delim("data/ASFIS_sp_2023.txt")
```

```
Rows: 13615 Columns: 14

-- Column specification --------------------------------------------------------
Delimiter: ","
chr (13): Taxonomic_Code, Alpha3_Code, Scientific_Name, English_name, French...
dbl  (1): ISSCAAP_Group

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
asfis_data
```

```
# A tibble: 13,615 x 14
   ISSCAAP_Group Taxonomic_Code Alpha3_Code Scientific_Name         English_name
           <dbl> <chr>          <chr>       <chr>                   <chr>
 1            11 140001000101   YCL         Cycleptus elongatus     Blue sucker
 2            11 140001000201   DEU         Deltistes luxatus       Lost River ~
 3            11 140001000401   ATC         Catostomus catostomus   Longnose su~
 4            11 140001000402   ATO         Catostomus commersoni   White sucker
 5            11 140001000403   ATS         Catostomus latipinnis   Flannelmout~
 6            11 140001000404   ATU         Catostomus macrocheilus Largescale ~
 7            11 140001000405   ATE         Catostomus tahoensis    Tahoe sucker
 8            11 140001000601   MOG         Moxostoma congestum     Gray redhor~
 9            11 140001000602   MOE         Moxostoma erythrurum    Golden redh~
10            11 140001000603   MOM         Moxostoma macrolepidot~ Shorthead r~
# i 13,605 more rows
# i 9 more variables: French_name <chr>, Spanish_name <chr>, Arabic_name <chr>,
#   Chinese_name <chr>, Russian_name <chr>, Author <chr>, Family <chr>,
#   Order <chr>, FishStat_Data <chr>
```

```r
names(asfis_data)
```

```
 [1] "ISSCAAP_Group"   "Taxonomic_Code"  "Alpha3_Code"     "Scientific_Name"
 [5] "English_name"    "French_name"     "Spanish_name"    "Arabic_name"
 [9] "Chinese_name"    "Russian_name"    "Author"          "Family"
[13] "Order"           "FishStat_Data"
```

```r
asfis_data <- janitor::clean_names(asfis_data)
asfis_data
```

```
# A tibble: 13,615 x 14
   isscaap_group taxonomic_code alpha3_code scientific_name         english_name
           <dbl> <chr>          <chr>       <chr>                   <chr>
 1            11 140001000101   YCL         Cycleptus elongatus     Blue sucker
 2            11 140001000201   DEU         Deltistes luxatus       Lost River ~
 3            11 140001000401   ATC         Catostomus catostomus   Longnose su~
 4            11 140001000402   ATO         Catostomus commersoni   White sucker
 5            11 140001000403   ATS         Catostomus latipinnis   Flannelmout~
 6            11 140001000404   ATU         Catostomus macrocheilus Largescale ~
 7            11 140001000405   ATE         Catostomus tahoensis    Tahoe sucker
 8            11 140001000601   MOG         Moxostoma congestum     Gray redhor~
 9            11 140001000602   MOE         Moxostoma erythrurum    Golden redh~
10            11 140001000603   MOM         Moxostoma macrolepidot~ Shorthead r~
# i 13,605 more rows
# i 9 more variables: french_name <chr>, spanish_name <chr>, arabic_name <chr>,
#   chinese_name <chr>, russian_name <chr>, author <chr>, family <chr>,
#   order <chr>, fish_stat_data <chr>
```

```r
# 1a. How many columns does the data have?
num_columns <- ncol(asfis_data)
print(num_columns)
```

```
[1] 14
```

```r
# 1b. How many rows does the data have?
num_rows <- nrow(asfis_data)
print(num_rows)
```

```
[1] 13615
```

```r
# 1c. Print the 10 most occurring families in the dataset
top_families <- asfis_data |>
  count(family) |>
  slice_max(n, n = 10)
print(top_families)
```

```
# A tibble: 10 x 2
   family          n
   <chr>       <int>
 1 CYPRINIDAE    512
 2 CICHLIDAE     315
 3 GOBIIDAE      304
 4 SERRANIDAE    236
 5 SCIAENIDAE    185
 6 CARANGIDAE    152
 7 CLUPEIDAE     141
 8 SCORPAENIDAE  136
 9 PENAEIDAE     134
10 VENERIDAE     132
```

```
# 1d. What are the 10 least occurring isscaap_group?
least_isscaap <- asfis_data  |>
  count(isscaap_group) |>
  arrange(n) |>
  slice_min(n, n = 10)
print(least_isscaap)
```

```
# A tibble: 10 x 2
   isscaap_group     n
           <dbl> <int>
 1            39     4
 2            64     9
 3            22    12
 4            61    16
 5            75    17
 6            72    20
 7            73    24
 8            81    24
 9            58    26
10            94    27
```

```
# 2. Group the data by the order column and count unique values
order_summary <- asfis_data  |>
  group_by(order)  |>
  summarize(
    unique_isscaap = n_distinct(isscaap_group),
    unique_taxonomic = n_distinct(taxonomic_code),
    unique_alpha3 = n_distinct(alpha3_code)
  )  |>
  arrange(desc(unique_alpha3))
print(order_summary)
```

```
# A tibble: 152 x 4
   order             unique_isscaap unique_taxonomic unique_alpha3
   <chr>                      <int>            <int>         <int>
 1 PERCOIDEI                      6             2370          2370
 2 BIVALVIA                       6              793           793
 3 SILURIFORMES                   2              658           658
 4 GASTROPODA                     3              646           646
 5 CYPRINIFORMES                  1              616           616
 6 NATANTIA                       2              491           491
 7 SCORPAENIFORMES                3              450           450
 8 GOBIOIDEI                      2              368           368
 9 PLEURONECTIFORMES              1              320           320
10 CEPHALOPODA                    1              317           317
# i 142 more rows
```

```
# 3. How many fish have French names?
num_french_names <- sum(!is.na(asfis_data$french_name))
print(num_french_names)
```

```
[1] 5754
```

```
# 4. How many fish have all common names?
all_names_count <- asfis_data  |>
  filter(!is.na(french_name) & !is.na(spanish_name) & !is.na(english_name))  |>
  nrow()
print(all_names_count)
```

```
[1] 4578
```

```
# 5. Create a new variable called order_lower
asfis_data <- asfis_data %>%
  mutate(order_lower = tolower(order))
head(asfis_data)
```

```
# A tibble: 6 x 15
  isscaap_group taxonomic_code alpha3_code scientific_name         english_name
          <dbl> <chr>          <chr>       <chr>                   <chr>
1            11 140001000101   YCL         Cycleptus elongatus     Blue sucker
2            11 140001000201   DEU         Deltistes luxatus       Lost River s~
3            11 140001000401   ATC         Catostomus catostomus   Longnose suc~
4            11 140001000402   ATO         Catostomus commersoni   White sucker
5            11 140001000403   ATS         Catostomus latipinnis   Flannelmouth~
6            11 140001000404   ATU         Catostomus macrocheilus Largescale s~
# i 10 more variables: french_name <chr>, spanish_name <chr>,
#   arabic_name <chr>, chinese_name <chr>, russian_name <chr>, author <chr>,
#   family <chr>, order <chr>, fish_stat_data <chr>, order_lower <chr>
```

```r
# 6. Filter the dataset for order_lower == "pelecaniformes" and non-missing spanish_name
filtered_data <- asfis_data  |>
  filter(order_lower == "pelecaniformes" & !is.na(spanish_name))
print(filtered_data)
```

```
# A tibble: 9 x 15
  isscaap_group taxonomic_code alpha3_code scientific_name          english_name
          <dbl> <chr>          <chr>       <chr>                    <chr>
1            NA 562001000501   TWH         Pelecanus thagus         Peruvian pe~
2            NA 562002000101   ISQ         Phalacrocorax atriceps   Imperial sh~
3            NA 562002000102   ISW         Phalacrocorax aristotel~ European sh~
4            NA 562002000103   ISY         Phalacrocorax carbo      Great cormo~
5            NA 5620020XXXXX   ITV         Phalacrocoracidae        Cormorants ~
6            NA 562003000301   MVR         Morus serrator           Australasia~
7            NA 562003000302   MVB         Morus bassanus           Northern ga~
8            NA 562003000303   MWE         Morus capensis           Cape gannet
9            NA 562003000601   DSQ         Sula dactylatra          Masked booby
# i 10 more variables: french_name <chr>, spanish_name <chr>,
#   arabic_name <chr>, chinese_name <chr>, russian_name <chr>, author <chr>,
#   family <chr>, order <chr>, fish_stat_data <chr>, order_lower <chr>
```

```r
# 7. Filter the dataset and group by family
filtered_family <- asfis_data  |>
  filter(order_lower %in% c("bryozoa", "squamate"))  |>
  group_by(family)  |>
  summarize(count = n())  |>
  filter(count > 1)  |>
  arrange(desc(count))
print(filtered_family)
```

```
# A tibble: 3 x 2
  family        count
  <chr>         <int>
1 FLUSTRIDAE        4
2 SMITTINIDAE       3
3 ALCYONIDIIDAE     2
```

```
# 8. Count the number of authors that gave scientific names
author_count <- asfis_data  |>
  count(scientific_name)  |>
  arrange(desc(n))
print(author_count)
```

```
# A tibble: 13,602 x 2
   scientific_name      n
   <chr>            <int>
 1 Actinopterygii       6
 2 Clupeoidei           2
 3 Crustacea            2
 4 Elasmobranchii       2
 5 Gobiidae             2
 6 Mollusca             2
 7 Palaemonidae         2
 8 Perciformes          2
 9 Testudinata          2
10 Aaptosyax grypus     1
# i 13,592 more rows
```

```
# 9. Count families with 100 or more occurrences
families_100_or_more <- asfis_data  |>
  count(family)  |>
  filter(n >= 100)
num_families_100_or_more <- nrow(families_100_or_more)
print(num_families_100_or_more)
```

```
[1] 18
```

The `echo: false` option disables the printing of code (only output is displayed).