Database Acquisition

Madie Munro

4/6/2023

# Introduction:

## File Reading:

*The following code chunk below imports in the needed libraries for table ripping:*

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.0 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.1 ✔ tibble 3.1.8  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts to become errors

library(dplyr)  
library(naniar)  
library(stringr)

*Next, the original dataset was converted from an Excel file to a CSV in order to read in the data easier (see 03\_incremental folder for the .csv data). In addition, a column lookup dataset was read in, containing the new names for the columns of the database; extra na values were dropped after table was read in:*

fish <- read\_csv("C:/Users/s69m456/Desktop/FishXing(HEC-RAS) Independent Study/R\_database\_work/01\_acquisition/03\_incremental/fish\_info.csv")

## New names:  
## Rows: 156 Columns: 40  
## ── Column specification  
## ──────────────────────────────────────────────────────── Delimiter: "," chr  
## (30): Common Name, Scientific Name, Swim Mode, Min Speed, Max Speed, Swi... dbl  
## (10): BD/TL, TL/SL, TL/FL, Default Time, B a, B b, B c, P a, P b, P c  
## 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.  
## • `Units` -> `Units...7`  
## • `Variability` -> `Variability...8`  
## • `Variability Type` -> `Variability Type...9`  
## • `Variability` -> `Variability...14`  
## • `Variability Type` -> `Variability Type...15`  
## • `Units` -> `Units...16`

fish\_cols <- read\_csv("C:/Users/s69m456/Desktop/FishXing(HEC-RAS) Independent Study/R\_database\_work/01\_acquisition/01\_input/column\_lookup.csv")

## Rows: 16384 Columns: 3  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (3): OldName, NewName, ColDesc  
##   
## 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.

fish\_cols <- drop\_na(fish\_cols)

## Column Renaming:

*The columns stored in “fish” were changed to reflect the new column names to be used in the updated database. This was accomplished by creating a subset of the “fish\_cols” data, pivoting the data wider in order for the new column names to be the data column names, and then renaming each column name in “fish” to the column names in the subsetted data. Lastly, the column “FishIDX” was added to the end of the “fish” dataset:*

cols\_sub <- fish\_cols[, c("NewName", "ColDesc")]  
fcols <- pivot\_wider(cols\_sub, names\_from = "NewName", values\_from = "ColDesc")  
  
colnames(fish) = colnames(fcols)  
fish$FishIDX <- 1:nrow(fish)  
  
head(fish)

## # A tibble: 6 × 41  
## CommonName Scient…¹ SwimM…² MinSp…³ MaxSp…⁴ SwimS…⁵ SwimU…⁶ SwimV…⁷ SwimV…⁸  
## <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>   
## 1 Twaite shad Alosa f… Prolon… NR 75 75 cm/s NR <NA>   
## 2 American shad Alosa s… Burst 244 142 193 cm/s NR <NA>   
## 3 American shad Alosa s… Burst 55 76 65.5 cm/s NR <NA>   
## 4 American shad Alosa s… Prolon… 92 213 152.5 cm/s NR <NA>   
## 5 American shad Alosa s… Prolon… 40 52 46 cm/s NR <NA>   
## 6 Stone loach Barbatu… Prolon… NR NR 60.8 cm/s NR <NA>   
## # … with 32 more variables: LengthType <chr>, MinLength <chr>, MaxLength <chr>,  
## # LengthValue <chr>, LengthVariabilityValue <chr>,  
## # LengthVariabilityType <chr>, LengthUnits <chr>, `BD/TLRatio` <dbl>,  
## # `TL/SLRatio` <dbl>, `TL/FLRatio` <dbl>, SwimTime <dbl>, MinTime <chr>,  
## # MaxTime <chr>, TimeBetweenIncrements <chr>, VelocityIncrements <chr>,  
## # SwimmingTimeAtMaxVelocity <chr>, `Leap(Y/N)` <chr>, Temperature <chr>,  
## # BurstCoeffA <dbl>, BurstCoeffB <dbl>, BurstCoeffC <dbl>, …

## Data Mutation:

*Some numeric column values gave a range of values rather than one set value, so they were split into two respective columns: low and high. The first column to be split as such was the one for water temerature:*

fish\_temps <-  
 fish$Temperature %>%  
 strsplit(split = "\u00B1|to") %>%  
 lapply(., function(x) {   
 x <-   
 as.vector(x)%>%  
 as.numeric   
 if(length(x) == 2){  
 return(x)  
 }  
 else {  
 x <- rep(x,2)  
 }  
 }) %>% do.call(rbind, .) %>%  
as.data.frame(.)  
  
fish\_temps\_low <- fish\_temps$V1  
fish\_temps\_low[8] <- (fish\_temps$V1[8] - fish\_temps$V2[8])  
fish\_temps\_low[113] <- (fish\_temps$V1[113] - fish\_temps$V2[113])  
fish\_temps\_low[114] <- (fish\_temps$V1[114] - fish\_temps$V2[114])  
  
fish\_temps\_high <- fish\_temps$V2  
fish\_temps\_high[8] <- (fish\_temps$V1[8] + fish\_temps$V2[8])  
fish\_temps\_high[113] <- (fish\_temps$V1[113] + fish\_temps$V2[113])  
fish\_temps\_high[114] <- (fish\_temps$V1[114] + fish\_temps$V2[114])

*Swim variability also got the same treatment:*

fish\_swim\_var\_low <-  
 fish$SwimVariabilityValue %>%  
 str\_match\_all("[0-9.]+") %>%   
 lapply(., function(x) {  
 x <-   
 as.vector(x)%>%  
 as.numeric   
 return(x[1])  
 })%>% unlist  
  
fish\_swim\_var\_high <-  
 fish$SwimVariabilityValue %>%  
 str\_match\_all("[0-9.]+") %>%   
 lapply(., function(x) {  
 x <-   
 as.vector(x)%>%  
 as.numeric  
 ifelse(x[1] == "numeric(0)", replace\_with\_na(replace = x[1]), return(x[length(x)]))  
 }) %>% unlist

*Once the more problematic columns were appropriately split apart, each of the columns with numeric values were converted to be as such (character-to-numeric). Some values were converted to double, some were just integer. The modified data was stored under “fish\_test”:*

fish\_test <- fish %>%  
 mutate\_at(c("MinSpeed", "MaxSpeed", "SwimSpeed", "SwimVariabilityValue", "MinLength", "MaxLength", "LengthValue", "LengthVariabilityValue",  
 "SwimTime", "MinTime", "MaxTime", "TimeBetweenIncrements", "VelocityIncrements", "SwimmingTimeAtMaxVelocity"), as.numeric) %>%  
 mutate(NumberOfFishTested = as.integer(NumberOfFishTested)) %>%  
 mutate(fish\_temps\_high = as.double(fish\_temps\_high)) %>%  
 mutate(fish\_temps\_low = as.double(fish\_temps\_low)) %>%  
 mutate(fish\_swim\_var\_high = as.double(fish\_swim\_var\_high)) %>%  
 mutate(fish\_swim\_var\_low = as.double(fish\_swim\_var\_low))  
  
head(fish\_test)

## # A tibble: 6 × 45  
## CommonName Scient…¹ SwimM…² MinSp…³ MaxSp…⁴ SwimS…⁵ SwimU…⁶ SwimV…⁷ SwimV…⁸  
## <chr> <chr> <chr> <dbl> <dbl> <dbl> <chr> <dbl> <chr>   
## 1 Twaite shad Alosa f… Prolon… NA 75 75 cm/s NA <NA>   
## 2 American shad Alosa s… Burst 244 142 193 cm/s NA <NA>   
## 3 American shad Alosa s… Burst 55 76 65.5 cm/s NA <NA>   
## 4 American shad Alosa s… Prolon… 92 213 152. cm/s NA <NA>   
## 5 American shad Alosa s… Prolon… 40 52 46 cm/s NA <NA>   
## 6 Stone loach Barbatu… Prolon… NA NA 60.8 cm/s NA <NA>   
## # … with 36 more variables: LengthType <chr>, MinLength <dbl>, MaxLength <dbl>,  
## # LengthValue <dbl>, LengthVariabilityValue <dbl>,  
## # LengthVariabilityType <chr>, LengthUnits <chr>, `BD/TLRatio` <dbl>,  
## # `TL/SLRatio` <dbl>, `TL/FLRatio` <dbl>, SwimTime <dbl>, MinTime <dbl>,  
## # MaxTime <dbl>, TimeBetweenIncrements <dbl>, VelocityIncrements <dbl>,  
## # SwimmingTimeAtMaxVelocity <dbl>, `Leap(Y/N)` <chr>, Temperature <chr>,  
## # BurstCoeffA <dbl>, BurstCoeffB <dbl>, BurstCoeffC <dbl>, …

*The last step taken before the full data was ripped into seperate tables for SQL querying was the renaming of the new columns (low/high temp/swim variability) and replacing the question marks in the “Leap(y/n)” column with “na”:*

colnames(fish\_test)[42] <- "WaterTempHigh"   
colnames(fish\_test)[43] <- "WaterTempLow"   
colnames(fish\_test)[44] <- "SwimVariabilityValueHigh"  
colnames(fish\_test)[45] <- "SwimVariabilityValueLow"  
  
fish\_test <- fish\_test %>% replace\_with\_na\_at(.vars = "Leap(Y/N)", condition = ~.x == "?")  
  
head(fish\_test)

## # A tibble: 6 × 45  
## CommonName Scient…¹ SwimM…² MinSp…³ MaxSp…⁴ SwimS…⁵ SwimU…⁶ SwimV…⁷ SwimV…⁸  
## <chr> <chr> <chr> <dbl> <dbl> <dbl> <chr> <dbl> <chr>   
## 1 Twaite shad Alosa f… Prolon… NA 75 75 cm/s NA <NA>   
## 2 American shad Alosa s… Burst 244 142 193 cm/s NA <NA>   
## 3 American shad Alosa s… Burst 55 76 65.5 cm/s NA <NA>   
## 4 American shad Alosa s… Prolon… 92 213 152. cm/s NA <NA>   
## 5 American shad Alosa s… Prolon… 40 52 46 cm/s NA <NA>   
## 6 Stone loach Barbatu… Prolon… NA NA 60.8 cm/s NA <NA>   
## # … with 36 more variables: LengthType <chr>, MinLength <dbl>, MaxLength <dbl>,  
## # LengthValue <dbl>, LengthVariabilityValue <dbl>,  
## # LengthVariabilityType <chr>, LengthUnits <chr>, `BD/TLRatio` <dbl>,  
## # `TL/SLRatio` <dbl>, `TL/FLRatio` <dbl>, SwimTime <dbl>, MinTime <dbl>,  
## # MaxTime <dbl>, TimeBetweenIncrements <dbl>, VelocityIncrements <dbl>,  
## # SwimmingTimeAtMaxVelocity <dbl>, `Leap(Y/N)` <chr>, Temperature <chr>,  
## # BurstCoeffA <dbl>, BurstCoeffB <dbl>, BurstCoeffC <dbl>, …

# Individual Table Creation: