Filter Data

Derek H. Ogle, Northland College

2-Mar-2015

Preliminaries

```
> library(fishWiDNR)   # for setDBClasses(), expandCounts()
> library(dplyr)       # for select(), filter()
> library(FSA)         # for Summarize()
```

```
> setwd("C:/aaaWork/Web/fishR/Courses/WiDNR_Statewide_2015/Day1_IntroR_FMData")
> d <- read.csv("FMDB_Sawyer.csv",stringsAsFactors=FALSE)
> d <- setDBClasses(d,type="RDNR")
> d <- expandCounts(d,~Number.of.Fish,~Length.or.Lower.Length.IN+Length.Upper.IN,new.name="Len")
Some rows (13884, 20543) had zero counts in Number.of.Fish.
17430 rows had an individual measurement.
3926 rows with multiple measurements were expanded to 36156 rows of individual measurements.
```

```
> names(d)
 [1] "County"                   "Waterbody.Name"           "WBIC"
 [4] "Survey.Year"              "Station.Name"             "Swims.Station.Id"
 [7] "Site.Seq.No"              "Srvy.Seq.No"              "Survey.Begin.Date"
[10] "Survey.End.Date"          "Survey.Status"            "Data.Entry.Name"
[13] "Entry.Date"               "Visit.Fish.Seq.No"        "Visit.Type"
[16] "Gear"                     "Sample.Date"              "Substation.Name"
[19] "Target.Species"           "Fish.Data.Seq.No"         "Net.Number"
[22] "Species.Code"             "Species"                  "Length.or.Lower.Length.IN"
[25] "Length.Upper.IN"          "Length.or.Lower.Length.MM" "Length.Upper.MM"
[28] "Weight.Pounds"            "Weight.Grams"             "Gender"
[31] "Disease"                  "Injury.Type"              "Age..observed.annuli."
[34] "Edge.Counted.Desc"        "Age.Structure"            "Mark.Given"
[37] "Mark.Found"               "Second.Mark.Found"        "Tag.Number.Given"
[40] "Second.Tag.Number.Given"  "Tag.Number.Found"         "Second.Tag.Number.Found"
[43] "YOY"                      "Entry.Date.1"             "Last.Update.Date"
[46] "Data.Ent.Name"            "Last.Update.Name"         "Invalid.Species"
[49] "Non.Standard.Bin"         "Length.Unit.Error"        "Length.Outside.Range"
[52] "Count.Outside.Range"      "Status.Code"              "Len"
[55] "lennote"
```

Selecting Variables – select()

```
> d1 <- select(d,Waterbody.Name,Gear,Survey.Year,Species,Len,Weight.Pounds,Gender,Mark.Given)
> head(d1)
   Waterbody.Name     Gear Survey.Year       Species  Len Weight.Pounds Gender Mark.Given
11   ASHEGON LAKE FYKE NET        2010 BLACK CRAPPIE  5.0            NA
12   ASHEGON LAKE FYKE NET        2010 BLACK CRAPPIE  9.8            NA
13   ASHEGON LAKE FYKE NET        2010 BLACK CRAPPIE 10.2            NA
14   ASHEGON LAKE FYKE NET        2010 BLACK CRAPPIE 10.3            NA
15   ASHEGON LAKE FYKE NET        2010 BLACK CRAPPIE 12.0            NA
16   ASHEGON LAKE FYKE NET        2010 BLACK CRAPPIE 12.2            NA


> tail(d1)
      Waterbody.Name     Gear Survey.Year Species  Len Weight.Pounds Gender Mark.Given
52724      TEAL LAKE FYKE NET        2010 WALLEYE 15.0            NA      M
52725      TEAL LAKE FYKE NET        2010 WALLEYE 15.3            NA      M
52726      TEAL LAKE FYKE NET        2010 WALLEYE 17.3            NA      M
52727      TEAL LAKE FYKE NET        2010 WALLEYE 17.0            NA      M
52728      TEAL LAKE FYKE NET        2010 WALLEYE 17.9            NA      M
52729      TEAL LAKE FYKE NET        2010 WALLEYE 17.8            NA      M


> tmp <- select(d,County:Swims.Station.Id)
> head(tmp)
   County Waterbody.Name    WBIC Survey.Year                       Station.Name Swims.Station.Id
11 SAWYER   ASHEGON LAKE 2448800        2010 ASHEGON LAKE_GENERAL LAKE STATION         10005674
12 SAWYER   ASHEGON LAKE 2448800        2010 ASHEGON LAKE_GENERAL LAKE STATION         10005674
13 SAWYER   ASHEGON LAKE 2448800        2010 ASHEGON LAKE_GENERAL LAKE STATION         10005674
14 SAWYER   ASHEGON LAKE 2448800        2010 ASHEGON LAKE_GENERAL LAKE STATION         10005674
15 SAWYER   ASHEGON LAKE 2448800        2010 ASHEGON LAKE_GENERAL LAKE STATION         10005674
16 SAWYER   ASHEGON LAKE 2448800        2010 ASHEGON LAKE_GENERAL LAKE STATION         10005674


> tmp <- select(d,-(Station.Name:Status.Code))
> head(tmp)
   County Waterbody.Name    WBIC Survey.Year  Len         lennote
11 SAWYER   ASHEGON LAKE 2448800        2010  5.0 Observed length
12 SAWYER   ASHEGON LAKE 2448800        2010  9.8 Observed length
13 SAWYER   ASHEGON LAKE 2448800        2010 10.2 Observed length
14 SAWYER   ASHEGON LAKE 2448800        2010 10.3 Observed length
15 SAWYER   ASHEGON LAKE 2448800        2010 12.0 Observed length
16 SAWYER   ASHEGON LAKE 2448800        2010 12.2 Observed length


> tmp <- select(d,starts_with("Length"))   # there is also an ends_with
> names(tmp)
[1] "Length.or.Lower.Length.IN" "Length.Upper.IN"           "Length.or.Lower.Length.MM"
[4] "Length.Upper.MM"           "Length.Unit.Error"         "Length.Outside.Range"


> tmp <- select(d,Srvy.Seq.No,Species,Len,contains("Mark"))
> head(tmp)
   Srvy.Seq.No       Species  Len Mark.Given Mark.Found Second.Mark.Found
11    56064296 BLACK CRAPPIE  5.0
12    56064296 BLACK CRAPPIE  9.8
13    56064296 BLACK CRAPPIE 10.2
14    56064296 BLACK CRAPPIE 10.3
15    56064296 BLACK CRAPPIE 12.0
16    56064296 BLACK CRAPPIE 12.2
```

Selecting Individuals – filter()

```
> levels(d1$Gear)
[1] "BACKPACK SHOCKER"  "BOOM SHOCKER"      "BOTTOM GILL NET"   "FYKE NET"
[5] "MINI BOOM SHOCKER" "STREAM SHOCKER"

> xtabs(~Gear,data=d1)
Gear
 BACKPACK SHOCKER      BOOM SHOCKER   BOTTOM GILL NET          FYKE NET MINI BOOM SHOCKER
             1049             18944               182             25177               386
   STREAM SHOCKER
             7850

> xtabs(~Waterbody.Name+Gear,data=d1)      # only partial results shown

                          Gear
Waterbody.Name             BACKPACK SHOCKER BOOM SHOCKER BOTTOM GILL NET FYKE NET
  ASHEGON LAKE                            0            0               0       98
  BADGER CREEK                          105            0               0        0
  BARBER LAKE                             0          661               0     2179
  BARKER LAKE                             0            0              19        0
  BILLY BOY FLOWAGE                       0            0               0      104
  BLACK DAN LAKE                          0          554               0      953
  BLAISDELL LAKE                          0            0               7        0
  BLUEBERRY LAKE                          0           61               0        0
  BRUNET RIVER                            0            0               0        0
  CHIPPANAZIE CREEK                      69            0               0        0
  CHIPPEWA RIVER                          0            0             140        0
  CONNORS LAKE                            0          735               0     2126
  DURPHEE LAKE                            0          693               0      386
  EAST FORK CHIPPEWA RIVER                0            0               0        0
  EDDY CREEK                              0            0               0        0

> tmp <- filter(d1,Waterbody.Name=="BARBER LAKE")
> xtabs(~Waterbody.Name,tmp)               # only partial results shown

Waterbody.Name
            ASHEGON LAKE             BADGER CREEK              BARBER LAKE              BARKER LAKE
                       0                        0                     2840                        0
       BILLY BOY FLOWAGE           BLACK DAN LAKE          BLAISDELL LAKE          BLUEBERRY LAKE
                       0                        0                        0                        0
            BRUNET RIVER        CHIPPANAZIE CREEK          CHIPPEWA RIVER            CONNORS LAKE
                       0                        0                        0                        0
            DURPHEE LAKE EAST FORK CHIPPEWA RIVER              EDDY CREEK          FLAMBEAU RIVER
                       0                        0                        0                        0

> tmp <- droplevels(tmp)
> xtabs(~Waterbody.Name,tmp)
Waterbody.Name
BARBER LAKE
       2840

> tmp <- filter(d1,Waterbody.Name %in% c("BARBER LAKE","LAKE CHETAC"))
> tmp <- droplevels(tmp)
> xtabs(~Waterbody.Name,tmp)
Waterbody.Name
BARBER LAKE LAKE CHETAC
       2840        6946
```

```
> LCblg <- filter(d1,Waterbody.Name=="LAKE CHETAC",Species=="BLUEGILL")
> xtabs(~Gear,LCblg)
Gear
 BACKPACK SHOCKER      BOOM SHOCKER   BOTTOM GILL NET          FYKE NET MINI BOOM SHOCKER
                0               398                 0               191                 0
   STREAM SHOCKER
                0

> LCblg <- filter(LCblg,Gear=="BOOM SHOCKER")
> Summarize(~Len,data=LCblg)
       n     mean       sd      min       Q1   median       Q3      max percZero
 398.000    5.984    1.163    3.000    5.000    6.200    6.900    8.900    0.000

> LCblgPREF <- filter(LCblg,Len>=7)
> Summarize(~Len,data=LCblgPREF)
       n     mean       sd      min       Q1   median       Q3      max percZero
 95.0000   7.3632   0.3461   7.0000   7.1000   7.3000   7.5000   8.9000   0.0000

> sturgWts <- filter(d1,Species=="LAKE STURGEON",!is.na(Weight.Pounds))
> head(sturgWts)
  Waterbody.Name            Gear Survey.Year       Species  Len Weight.Pounds Gender Mark.Given
1    BARKER LAKE BOTTOM GILL NET        2010 LAKE STURGEON 58.0          43.9      U        PIT
2    BARKER LAKE BOTTOM GILL NET        2010 LAKE STURGEON 61.5          70.5      U        PIT
3    BARKER LAKE BOTTOM GILL NET        2010 LAKE STURGEON 59.7          55.6      U        PIT
4    BARKER LAKE BOTTOM GILL NET        2010 LAKE STURGEON 62.5          66.5      U
5    BARKER LAKE BOTTOM GILL NET        2010 LAKE STURGEON 55.7          38.8      U
6    BARKER LAKE BOTTOM GILL NET        2010 LAKE STURGEON 56.4          45.7      U        PIT
```

Application Assignment

Create a script that performs the following tasks:

1. Load and prepare (set classes, expand counts, examine structure) your FM data in R (**HINT:** *use all or some of your script from the first application assignment*). Call this the *original data.frame*.
2. Create a data.frame that removes all variables related to the database (e.g., when data was entered, who entered it, error flags, etc.).
3. Examine the sample size per water body and gear combination in the original data.frame.
4. Isolate (from the original data.frame) a water body of your choice and show the number of each species captured (in all gears).
5. Isolate (from the original data.frame) three water bodies of your choice and make one table that shows the number of each species captured in each water body.
6. Isolate (from the original data.frame) one species of fish from one gear used in one waterbody.
   - Construct a table of frequency of each sex.
   - Summarize the length variable.
7. (*Time Permitting*) Suppose the waterbody and species you chose above has a minimum length limit (make up the minimum length). Isolate those fish that would be legal. Show that your filtering was successful.
8. (*Time Permitting*) Repeat the previous questions but for a protected slot.
9. (*Time Permitting*) Repeat the previous questions but for a harvest slot.
10. (*Time Permitting*) List all water bodies and species for which a weight in pounds was recorded (begin with the original data.frame).

Save your script!