

Filtering Data

Derek H. Ogle, Northland College

4-Mar-2015

Preliminaries

```
> # clears objects in R workspace
> rm(list = ls())

> # load needed packages
> library(fishWiDNR) # for setDBClasses()
> library(dplyr)     # for select(), filter()
> library(FSA)       # for expandCounts(), Summarize(), filterD()

> # load FM data and expand lengths ... copied code from first handout
> setwd("C:/aaaWork/Web/fishR/Courses/WiDNR_Statewide_2015/Day1_IntroR_FMDData")
> d <- read.csv("SAWYER_fish_raw_data_012915.csv", stringsAsFactors=FALSE, na.strings=c("-", "NA", ""))
> d <- setDBClasses(d, type="RDNR")
> d <- expandCounts(d, ~Number.of.Fish, ~Length.or.Lower.Length.IN+Length.Upper.IN, new.name="Len")
> names(d)
[1] "County" "Waterbody.Name" "WBIC"
[4] "Survey.Year" "Station.Name" "Swims.Station.Id"
[7] "Site.Seq.No" "Survey.Seq.No" "Survey.Begin.Date"
[10] "Survey.End.Date" "Survey.Status" "Data.Entry.Name"
[13] "Visit.Fish.Seq.No" "Visit.Type" "Gear"
[16] "Sample.Date" "Substation.Name" "Target.Species"
[19] "Fish.Data.Seq.No" "Net.Number" "Species.Code"
[22] "Species" "Length.or.Lower.Length.IN" "Length.Upper.IN"
[25] "Length.or.Lower.Length.MM" "Length.Upper.MM" "Weight.Pounds"
[28] "Weight.Grams" "Gender" "Disease."
[31] "Injury.Type" "Age..observed.annuli." "Edge.Counted.Desc"
[34] "Age.Structure" "Mark.Given" "Mark.Found"
[37] "Second.Mark.Found" "Tag.Number.Given" "Second.Tag.Number.Given"
[40] "Tag.Number.Found" "Second.Tag.Number.Found" "YOY"
[43] "Entry.Date" "Last.Update.Date" "Data.Ent.Name"
[46] "Last.Update.Name" "Invalid.Species" "Non.Standard.Bin"
[49] "Length.Unit.Error" "Length.Outside.Range" "Count.Outside.Range"
[52] "Status.Code" "Len" "lennote"
```

Selecting Variables – select()

```
> d1 <- select(d,Waterbody.Name,Gear,Survey.Year,Species,Len,Weight.Pounds,Gender)
> headtail(d1)
```

	Waterbody.Name	Gear	Survey.Year	Species	Len	Weight.Pounds	Gender
19	SISSABAGAMA LAKE	FYKE NET	2010	YELLOW PERCH	8.2	NA	F
20	SISSABAGAMA LAKE	FYKE NET	2010	YELLOW PERCH	8.1	NA	F
21	SISSABAGAMA LAKE	FYKE NET	2010	YELLOW PERCH	8.6	NA	F
133236	WINDIGO LAKE BOOM	SHOCKER	2014	LARGEMOUTH BASS	12.0	NA	<NA>
133237	WINDIGO LAKE BOOM	SHOCKER	2014	WALLEYE	18.1	NA	<NA>
133238	WINDIGO LAKE BOOM	SHOCKER	2014	WALLEYE	18.4	NA	<NA>

```
> tmp <- select(d,County:Station.Name)
> headtail(tmp)
```

	County	Waterbody.Name	WBIC	Survey.Year	Station.Name
19	SAWYER	SISSABAGAMA LAKE	2393500	2010	SISSABAGAMA LAKE_GENERAL LAKE STATION
20	SAWYER	SISSABAGAMA LAKE	2393500	2010	SISSABAGAMA LAKE_GENERAL LAKE STATION
21	SAWYER	SISSABAGAMA LAKE	2393500	2010	SISSABAGAMA LAKE_GENERAL LAKE STATION
133236	SAWYER	WINDIGO LAKE	2046600	2014	WINDIGO LAKE_GENERAL LAKE STATION
133237	SAWYER	WINDIGO LAKE	2046600	2014	WINDIGO LAKE_GENERAL LAKE STATION
133238	SAWYER	WINDIGO LAKE	2046600	2014	WINDIGO LAKE_GENERAL LAKE STATION

```
> tmp <- select(d,-(Station.Name:Status.Code))
> headtail(tmp)
```

	County	Waterbody.Name	WBIC	Survey.Year	Len	lennote
19	SAWYER	SISSABAGAMA LAKE	2393500	2010	8.2	Observed length
20	SAWYER	SISSABAGAMA LAKE	2393500	2010	8.1	Observed length
21	SAWYER	SISSABAGAMA LAKE	2393500	2010	8.6	Observed length
133236	SAWYER	WINDIGO LAKE	2046600	2014	12.0	Expanded length
133237	SAWYER	WINDIGO LAKE	2046600	2014	18.1	Expanded length
133238	SAWYER	WINDIGO LAKE	2046600	2014	18.4	Expanded 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,Survey.Seq.No,Species,Len,contains("Mark"))
> headtail(tmp)
```

	Survey.Seq.No	Species	Len	Mark.Given	Mark.Found	Second.Mark.Found
19	39508941	YELLOW PERCH	8.2	<NA>	<NA>	<NA>
20	39508941	YELLOW PERCH	8.1	<NA>	<NA>	<NA>
21	39508941	YELLOW PERCH	8.6	<NA>	<NA>	<NA>
133236	515077184	LARGEMOUTH BASS	12.0	<NA>	<NA>	<NA>
133237	515077184	WALLEYE	18.1	<NA>	<NA>	<NA>
133238	515077184	WALLEYE	18.4	<NA>	<NA>	<NA>

Selecting Individuals – filter()

```
> levels(d1$Waterbody.Name)
[1] "ALDER CREEK"
[3] "BADGER CREEK"
[5] "BARKER LAKE"
[7] "BILLY BOY FLOWAGE"
[9] "BLAISDELL LAKE"
[11] "BLUEBERRY LAKE"
[13] "CALLAHAN LAKE"
[15] "CHIPPEWA RIVER"
[17] "COUDERAY RIVER"
[19] "DURPHEE LAKE"
[21] "EDDY CREEK"
[23] "FLAMBEAU RIVER"
[25] "GREEN LAKE"
[27] "GRINDSTONE LAKE"
[29] "HATCHERY CREEK"
[31] "HUNTER LAKE"
[33] "LAC COURTE OREILLES"
[35] "LAKE CHIPPEWA"
[37] "LITTLE LAC COURTE OREILLES"
[39] "LITTLE WEIRGOR CREEK"
[41] "LORETTA LAKE (U BRUNET FLOWAGE)"
[43] "LOWER CLAM LAKE"
[45] "MAPLE CREEK"
[47] "MOOSE LAKE"
[49] "MOSQUITO BROOK"
[51] "NAMEKAGON RIVER"
[53] "NO OFFICIAL WATERBODY NAME"
[55] "OSPREY CREEK"
[57] "PARTRIDGE CROP LAKE"
[59] "PINE CREEK"
[61] "ROUND LAKE"
[63] "SILVERTHORN LAKE"
[65] "SMITH LAKE"
[67] "SPIDER LAKE"
[69] "SWAN CREEK"
[71] "TEAL LAKE"
[73] "TIGER CAT FLOWAGE"
[75] "TOTAGATIC RIVER"
[77] "UNNAMED SINGLE-LINE STREAM T38N-R3W-S7"
[79] "UNNAMED SINGLE-LINE STREAM T41N-R5W-S22"
[81] "UPPER HOLLY LAKE"
[83] "WHITEFISH LAKE"
[85] "WINDIGO LAKE"

"ASHEGON LAKE"
"BARBER LAKE"
"BENSON CREEK"
"BLACK DAN LAKE"
"BLUEBERRY CREEK"
"BRUNET RIVER"
"CHIPPANAZIE CREEK"
"CONNORS LAKE"
"DEER LAKE"
"EAST FORK CHIPPEWA RIVER"
"EVERGREEN LAKE"
"FORTYONE CREEK"
"GRINDSTONE CREEK"
"HACKETT CREEK"
"HAYWARD LAKE"
"ISLAND LAKE"
"LAKE CHETAC"
"LAKE OF THE PINES"
"LITTLE ROUND LAKE"
"LOG CREEK"
"LOST LAND LAKE"
"LOWER HOLLY LAKE"
"MASON LAKE"
"MOOSE RIVER"
"MUD LAKE"
"NELSON LAKE"
"NORTH BRANCH TUPPER CREEK"
"OSPREY LAKE"
"PELICAN LAKE"
"RADISSON FLOWAGE"
"SAND LAKE"
"SISSABAGAMA LAKE"
"SMITH LAKE CREEK"
"SPRING LAKE"
"SWIFT CREEK"
"THORNAPPLE RIVER"
"TOTAGATIC FLOWAGE"
"TUPPER CREEK"
"UNNAMED SINGLE-LINE STREAM T40N-R4W-S24"
"UNNAMED SINGLE-LINE STREAM T41N-R9W-S32"
"VENISON CREEK"
"WINDFALL LAKE"
"WINTER LAKE (PRICE FLOWAGE)"
```

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

Waterbody.Name				
ALDER CREEK	ASHEGON LAKE	BADGER CREEK	BARBER LAKE	BARKER LAKE
139	98	392	2895	25
BENSON CREEK	BILLY BOY FLOWAGE	BLACK DAN LAKE	BLAISDELL LAKE	BLUEBERRY CREEK
74	104	1547	63	52
BLUEBERRY LAKE	BRUNET RIVER	CALLAHAN LAKE	CHIPPANAZIE CREEK	CHIPPEWA RIVER
876	2080	269	69	337

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

Waterbody.Name	Gear	BACKPACK	SHOCKER	BOOM	SHOCKER	BOTTOM	GILL	NET	FYKE	NET
ALDER CREEK			139		0			0		0
ASHEGON LAKE			0		0			0		98
BADGER CREEK			105		0			0		0
BARBER LAKE			0		716			0		2179
BARKER LAKE			0		0			25		0
BENSON CREEK			74		0			0		0
BILLY BOY FLOWAGE			0		0			0		104
BLACK DAN LAKE			0		594			0		953
BLAISDELL LAKE			0		22			41		0
BLUEBERRY CREEK			52		0			0		0
BLUEBERRY LAKE			0		706			0		170
BRUNET RIVER			133		0			0		0

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

Waterbody.Name	ALDER CREEK	ASHEGON LAKE	BADGER CREEK	BARBER LAKE	BARKER LAKE
	0	0	0	2895	0
BENSON CREEK	BILLY BOY FLOWAGE	BLACK DAN LAKE	BLAISDELL LAKE	BLUEBERRY CREEK	
	0	0	0	0	0
BLUEBERRY LAKE	BRUNET RIVER	CALLAHAN LAKE	CHIPPANAZIE CREEK	CHIPPEWA RIVER	
	0	0	0	0	0

```
> tmp <- droplevels(tmp)
> xtabs(~Waterbody.Name,data=tmp)
```

Waterbody.Name
BARBER LAKE
2895

```
> tmp <- filterD(d1,Waterbody.Name=="BARBER LAKE")
> xtabs(~Waterbody.Name,data=tmp)
```

Waterbody.Name
BARBER LAKE
2895

```
> tmp <- filterD(d1,Waterbody.Name %in% c("BARBER LAKE","LAKE CHETAC"))
> xtabs(~Waterbody.Name,data=tmp)
```

Waterbody.Name
BARBER LAKE LAKE CHETAC
2895 9182

```
> LCblg <- filterD(d1,Waterbody.Name=="LAKE CHETAC",Species=="BLUEGILL",Gear=="BOOM SHOCKER")
> xtabs(~Gear+Species,data=LCblg)
```

Gear	Species
BOOM SHOCKER	BLUEGILL
	740

```
> weird <- filterD(d1,Species=="Iowa Darter" | Weight.Pounds>100)
> weird
```

	Waterbody.Name	Gear	Survey.Year	Species	Len	Weight.Pounds	Gender
1	BLAISDELL LAKE	BOTTOM GILL NET	2012	LAKE STURGEON	72.7	100.80	<NA>
2	BLAISDELL LAKE	BOTTOM GILL NET	2013	LAKE STURGEON	67.1	105.16	U

```

> ( weird <- filterD(d1,Species=="IOWA DARTER" | Weight.Pounds>100) )
  Waterbody.Name      Gear Survey.Year      Species  Len Weight.Pounds Gender
1  BLAISDELL LAKE BOTTOM GILL NET      2012 LAKE STURGEON 72.7      100.80  <NA>
2  BLAISDELL LAKE BOTTOM GILL NET      2013 LAKE STURGEON 67.1      105.16    U
3  GRINDSTONE CREEK  STREAM SHOCKER      2010  IOWA DARTER  NA      NA  <NA>
4  GRINDSTONE CREEK  STREAM SHOCKER      2010  IOWA DARTER  NA      NA  <NA>
5  GRINDSTONE CREEK  STREAM SHOCKER      2010  IOWA DARTER  NA      NA  <NA>
6  GRINDSTONE CREEK  STREAM SHOCKER      2010  IOWA DARTER  NA      NA  <NA>
7  GRINDSTONE CREEK  STREAM SHOCKER      2010  IOWA DARTER  NA      NA  <NA>

> LCblgPREF <- filterD(LCblg,Len>=7)
> Summarize(~Len,data=LCblgPREF,digits=2)
      n      mean      sd      min      Q1      median      Q3      max percZero
154.00    7.34    0.33    7.00    7.10    7.30    7.50    8.90    0.00

> sturgWts <- filterD(d1,Species=="LAKE STURGEON",!is.na(Weight.Pounds))
> headtail(sturgWts)
  Waterbody.Name      Gear Survey.Year      Species  Len Weight.Pounds Gender
1    BARKER LAKE BOTTOM GILL NET      2010 LAKE STURGEON 58.0      43.9    U
2    BARKER LAKE BOTTOM GILL NET      2010 LAKE STURGEON 61.5      70.5    U
3    BARKER LAKE BOTTOM GILL NET      2010 LAKE STURGEON 59.7      55.6    U
247  BARKER LAKE BOTTOM GILL NET      2012 LAKE STURGEON 60.9      50.6  <NA>
248  BARKER LAKE BOTTOM GILL NET      2012 LAKE STURGEON 58.3      34.2  <NA>
249  BARKER LAKE BOTTOM GILL NET      2012 LAKE STURGEON 58.3      34.2  <NA>

```

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 datum 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 (regardless of gear).
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 question but for a protected slot.
9. (*Time Permitting*) Repeat the previous question 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!