data.frames I

Preliminaries

Load Necessary Packages

```
> library(FSA)     # for headtail()
> library(readxl)  # for read_excel()
> library(dplyr)   # for mutate()
```

Set Working Directory

```
> # Set your working directory to where your external data files (and scripts) are located.
> setwd("C:/aaaWork/Web/GitHub/RcourseNunavut2016/Handouts")
```

Load Data from External CSV File

```
> dSC <- read.csv("SawyerCo_reduced.csv")
> str(dSC)
'data.frame':   42682 obs. of  11 variables:
 $ waterbody: Factor w/ 11 levels "BLACK DAN LAKE",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ year     : int  2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 ...
 $ mon      : Factor w/ 7 levels "Apr","Aug","Jul",..: 7 7 7 7 7 7 7 7 7 7 ...
 $ gear     : Factor w/ 7 levels "BACKPACK SHOCKER",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ species  : Factor w/ 15 levels "Black Crappie",..: 13 13 13 13 13 13 13 13 13 13 ...
 $ len      : int  191 196 198 211 218 251 277 312 208 208 ...
 $ weight   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ sex      : Factor w/ 4 levels "","F","M","U": 1 1 1 1 1 1 1 1 1 1 ...
 $ age      : int  NA NA NA NA NA NA NA NA NA NA ...
 $ age_strux: Factor w/ 4 levels "","OTOLITH","SCALE",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ lennote  : Factor w/ 2 levels "Expanded length",..: 2 2 2 2 2 2 2 2 2 2 ...
```

```
> headtail(dSC)
          waterbody year mon         gear species len weight sex age age_strux         lennote
1     BLACK DAN LAKE 2012 Sep BOOM SHOCKER Walleye 191     NA      NA           Observed length
2     BLACK DAN LAKE 2012 Sep BOOM SHOCKER Walleye 196     NA      NA           Observed length
3     BLACK DAN LAKE 2012 Sep BOOM SHOCKER Walleye 198     NA      NA           Observed length
42680      SAND LAKE 2013 Oct BOOM SHOCKER Walleye 356     NA      NA           Observed length
42681      SAND LAKE 2013 Oct BOOM SHOCKER Walleye 356     NA      NA           Observed length
42682      SAND LAKE 2013 Oct BOOM SHOCKER Walleye 356     NA      NA           Observed length
```

```
> dSC$len
  [1] 191 196 198 211 218 251 277 312 208 208 264 264 173 188 168 178 208 165 178 183 185 160 165 165
 [25] 234 198 168 178 165 165 173 124 196 178 165 165 175 168 124 196 229 394 180 178 163 175 157 124
 [49] 155 175 267 163 124 244 180 183 168 180 183 117 282 173 173 178 229 213 104 173 178 163 211 170
 [ reached getOption("max.print") -- omitted 42610 entries ]
```

```
> dSC$len[1]
[1] 191


> dSC$len[c(1,3,5)]
[1] 191 198 218
```

Load Data from External Excel File

Demonstrate A Mess

```
> tmp <- read_excel("PG027.SA.Data.xlsx")
> str(tmp,list.len=10)              # list.len only used to save space
Classes 'tbl_df', 'tbl' and 'data.frame':   2045 obs. of  54 variables:
 $ Index                            : num  1101 1102 1103 1104 1105 ...
 $ Location - Name                  : chr  "Iqalujjuaq" "Iqalujjuaq" "Iqalujjuaq"
 $ AKA                              : chr  "Iqalugaarjuit Lake South" "Iqalugaarj
 $ Location - WB Code               : chr  "PG027" "PG027" "PG027" "PG027" ...
 $ Study Year                       : num  2012 2012 2012 2012 2012 ...
 $ Freshwater/ Seawater             : chr  "Freshwater" "Freshwater" "Freshwater"
 $ Data Collected by:               : chr  "DFO" "DFO" "DFO" "DFO" ...
 $ Station #                        : chr  "02A" "02A" "02A" "02A" ...
 $ Lift                             : chr  "nd" "nd" "nd" "nd" ...
 $ Net Type                         : chr  "multi" "multi" "multi" "multi" ...
  [list output truncated]


> tmp$Index[1:10]                   # positions used simply to limit output length
 [1] 1101 1102 1103 1104 1105 1106 1107 1108 1109 1110


> tmp$Location - Name[1:10]
Error: Unknown column 'Location'


> tmp$'Location - Name'[1:10]
 [1] "Iqalujjuaq" "Iqalujjuaq" "Iqalujjuaq" "Iqalujjuaq" "Iqalujjuaq" "Iqalujjuaq" "Iqalujjuaq"
 [8] "Iqalujjuaq" "Iqalujjuaq" "Iqalujjuaq"
```

An Alternative – More Work, But But More Useful Result

```
> # Get new names and defined data types
> ( meta <- read.csv("NU_metadata.csv",stringsAsFactors=FALSE) )
```

	old_names	new_names	new_types
1	Index	index	blank
2	Location - Name	loc	text
3	AKA	locAKA	text
4	Location - WB Code	locWB	text
5	Study Year	year	numeric
6	Freshwater/ Seawater	water.type	text
7	Data Collected by:	collector	text
8	Station #	station	text
9	Lift	lift	numeric
10	Net Type	net.type	text
11	Mesh Size (mm)	mesh.mm	numeric
12	Mesh Size\n\n(inch)	mesh.in	numeric
13	Species	spec	text
14	Sample #	sample	numeric
15	Fork Length (mm - fresh)	FL	numeric
16	Fork Length (mm - thawed)	FL.thawed	numeric
17	Round Weight (g - fresh)	wt	numeric
18	Round Weight (g - thawed)	wt.thawed	numeric
19	Dressed Weight (g)	wt.dressed	numeric
20	Sex (Male/ Female)	sex	text
21	Maturity	mat	text
22	Gonad Weight (g)	gonad.wt	numeric
23	Gonads Preserved\n\n(yes or no)	gonad.prsrvd	text
24	Otoliths (0/1/2)	oto.num	numeric
25	FIN CLIP (base for age)	finclip	text
26	DNA Tissue (fin tip)	dnatissue	text
27	Stomach Contents	stomach.contents	text
28	Stomach Preserved\n\n(yes or no)	stomach.prsrved	text
29	Muscle Tissue Frozen\n\n(yes or no)	muscle.frozen	text
30	Gill Arch Frozen\n\n(yes or no)	gillarch.frozen	text
31	Life History Type Suspected\n\n(Anadromous, Resident, Unknown)	life.hist	text
32	Use \n\n(Frozen / P / Released)	use	text
33	Age (Otolith)	age	numeric
34	Fecundity	fecundity	numeric
35	Average Egg Diameter (mm)	egg.diam	numeric
36	Latitude\n\n(dd.dddd)	lat	numeric
37	Longitude\n\n(dd.dddd)	long	numeric
38	Water Depth A (m)	depthA	numeric
39	Water Depth B (m)	depthB	numeric
40	Water Depth (m)	depth	numeric
41	Net Length (m)	net.len	numeric
42	Net Height (m)	net.height	numeric
43	Wind Direction	wind.dir	text
44	Wind Speed	wind.spd	text
45	Water Temp (oC)	temp.water	numeric
46	AIR Temp (oC)	temp.air	numeric
47	Sky	sky	text
48	Net Set Time	netset.time	date
49	Net Set Date	netset.date	date
50	Net Lift Time	netlift.time	date
51	Net Lift Date	netlift.date	date
52	Lake Zone	lake.zone	text
53	Tag #	tag	text
54	Remarks	remarks	text

```
> # Now read the data
> dNU <- read_excel("PG027.SA.Data.xlsx",na="nd",skip=1,
                     col_names=meta$new_names,col_types=meta$new_types)
> str(dNU)
Classes 'tbl_df', 'tbl' and 'data.frame':    2045 obs. of   53 variables:
 $ loc            : chr  "Iqalujjuaq" "Iqalujjuaq" "Iqalujjuaq" "Iqalujjuaq" ...
 $ locAKA         : chr  "Iqalugaarjuit Lake South" "Iqalugaarjuit Lake South" "Iqalugaarjuit Lake South" "Iq
 $ locWB          : chr  "PG027" "PG027" "PG027" "PG027" ...
 $ year           : num  2012 2012 2012 2012 2012 ...
 $ water.type     : chr  "Freshwater" "Freshwater" "Freshwater" "Freshwater" ...
 $ collector      : chr  "DFO" "DFO" "DFO" "DFO" ...
 $ station        : chr  "02A" "02A" "02A" "02A" ...
 $ lift           : num  NA NA NA NA NA NA NA NA NA NA ...
 $ net.type       : chr  "multi" "multi" "multi" "multi" ...
 $ mesh.mm        : num  63.5 114.3 114.3 63.5 38.1 ...
 $ mesh.in        : num  2.5 4.5 4.5 2.5 1.5 4.5 2.5 3.5 3.5 4.5 ...
 $ spec           : chr  "ARCH" "ARCH" "ARCH" "ARCH" ...
 $ sample         : num  20 15 16 21 11 14 19 13 12 17 ...
 $ FL             : num  421 510 638 399 705 646 476 520 579 507 ...
 $ FL.thawed      : num  NA NA NA NA NA NA NA NA NA NA ...
 $ wt             : num  729 1248 2832 485 3584 ...
 $ wt.thawed      : num  NA NA NA NA NA NA NA NA NA NA ...
 $ wt.dressed     : num  NA NA NA NA NA NA NA NA NA NA ...
 $ sex            : chr  "M" "M" "M" "F" ...
 $ mat            : chr  "R" "R" "R" "R" ...
 $ gonad.wt       : num  0.5 1.5 2.5 4.5 5.5 7.5 7.5 9.5 11.5 11.5 ...
 $ gonad.prsrvd   : chr  "N" "N" "N" "N" ...
 $ oto.num        : num  2 2 2 2 2 2 2 2 2 2 ...
 $ finclip        : chr  "Y" "Y" "Y" "Y" ...
 $ dnatissue      : chr  "Y" "Y" "Y" "Y" ...
 $ stomach.contents: chr  NA NA NA NA ...
 $ stomach.prsrved : chr  "N" "N" "N" "N" ...
 $ muscle.frozen  : chr  "Y" "Y" "Y" "Y" ...
 $ gillarch.frozen: chr  "Y" NA NA "Y" ...
 $ life.hist      : chr  NA NA NA NA ...
 $ use            : chr  "P" "P" "P" "P" ...
 $ age            : num  NA NA NA NA NA NA NA NA NA NA ...
 $ fecundity      : num  NA NA NA NA NA NA NA NA NA NA ...
 $ egg.diam       : num  NA NA NA NA NA NA NA NA NA NA ...
 $ lat            : num  65.7 65.7 65.7 65.7 65.7 ...
 $ long           : num  -64.8 -64.8 -64.8 -64.8 -64.8 ...
 $ depthA         : num  NA NA NA NA NA NA NA NA NA NA ...
 $ depthB         : num  NA NA NA NA NA NA NA NA NA NA ...
 $ depth          : num  NA NA NA NA NA NA NA NA NA NA ...
 $ net.len        : num  50 50 50 50 50 50 50 50 50 50 ...
 $ net.height     : num  1.83 1.83 1.83 1.83 1.83 1.83 1.83 1.83 1.83 1.83 ...
 $ wind.dir       : chr  NA NA NA NA ...
 $ wind.spd       : chr  "med" "med" "med" "med" ...
 $ temp.water     : num  NA NA NA NA NA NA NA NA NA NA ...
 $ temp.air       : num  NA NA NA NA NA NA NA NA NA NA ...
 $ sky            : chr  "overcast" "overcast" "overcast" "overcast" ...
 $ netset.time    : POSIXct, format: "1899-12-30 17:38:00" "1899-12-30 17:38:00" "1899-12-30 17:38:00" ...
 $ netset.date    : POSIXct, format: "2012-03-03" "2012-03-03" "2012-03-03" ...
 $ netlift.time   : POSIXct, format: "1899-12-30 21:13:00" "1899-12-30 21:13:00" "1899-12-30 21:13:00" ...
 $ netlift.date   : POSIXct, format: "2012-03-03" "2012-03-03" "2012-03-03" ...
 $ lake.zone      : chr  "benthic" "benthic" "benthic" "benthic" ...
 $ tag            : chr  NA NA NA NA ...
 $ remarks        : chr  NA "stomach parasites" NA NA ...
```

```
> # Adjust types of some variables
> dNU <- mutate(dNU,netset.time=format(netset.time,"%T"),netlift.time=format(netlift.time,"%T"),
+               fyear=factor(year),loc=factor(loc),locAKA=factor(locAKA),water.type=factor(water.type),
+               spec=factor(spec),sex=factor(sex),mat=factor(mat),life.hist=factor(life.hist))
> dNU <- as.data.frame(dNU)
> str(dNU)
'data.frame':    2045 obs. of  54 variables:
 $ loc             : Factor w/ 1 level "Iqalujjuaq": 1 1 1 1 1 1 1 1 1 1 ...
 $ locAKA          : Factor w/ 1 level "Iqalugaarjuit Lake South": 1 1 1 1 1 1 1 1 1 1 ...
 $ locWB           : chr  "PG027" "PG027" "PG027" "PG027" ...
 $ year            : num  2012 2012 2012 2012 2012 ...
 $ water.type      : Factor w/ 2 levels "Freshwater","seawater": 1 1 1 1 1 1 1 1 1 1 ...
 $ collector       : chr  "DFO" "DFO" "DFO" "DFO" ...
 $ station         : chr  "02A" "02A" "02A" "02A" ...
 $ lift            : num  NA NA NA NA NA NA NA NA NA NA ...
 $ net.type        : chr  "multi" "multi" "multi" "multi" ...
 $ mesh.mm         : num  63.5 114.3 114.3 63.5 38.1 ...
 $ mesh.in         : num  2.5 4.5 4.5 2.5 1.5 4.5 2.5 3.5 3.5 4.5 ...
 $ spec            : Factor w/ 1 level "ARCH": 1 1 1 1 1 1 1 1 1 1 ...
 $ sample          : num  20 15 16 21 11 14 19 13 12 17 ...
 $ FL              : num  421 510 638 399 705 646 476 520 579 507 ...
 $ FL.thawed       : num  NA NA NA NA NA NA NA NA NA NA ...
 $ wt              : num  729 1248 2832 485 3584 ...
 $ wt.thawed       : num  NA NA NA NA NA NA NA NA NA NA ...
 $ wt.dressed      : num  NA NA NA NA NA NA NA NA NA NA ...
 $ sex             : Factor w/ 3 levels "F","M","U": 2 2 2 1 2 2 1 1 1 1 ...
 $ mat             : Factor w/ 6 levels "I","M","R","RR",..: 3 3 3 3 3 3 3 3 5 3 ...
 $ gonad.wt        : num  0.5 1.5 2.5 4.5 5.5 7.5 7.5 9.5 11.5 11.5 ...
 $ gonad.prsrvd    : chr  "N" "N" "N" "N" ...
 $ oto.num         : num  2 2 2 2 2 2 2 2 2 2 ...
 $ finclip         : chr  "Y" "Y" "Y" "Y" ...
 $ dnatissue       : chr  "Y" "Y" "Y" "Y" ...
 $ stomach.contents: chr  NA NA NA NA ...
 $ stomach.prsrved : chr  "N" "N" "N" "N" ...
 $ muscle.frozen   : chr  "Y" "Y" "Y" "Y" ...
 $ gillarch.frozen : chr  "Y" NA NA "Y" ...
 $ life.hist       : Factor w/ 0 levels: NA NA NA NA NA NA NA NA NA NA ...
 $ use             : chr  "P" "P" "P" "P" ...
 $ age             : num  NA NA NA NA NA NA NA NA NA NA ...
 $ fecundity       : num  NA NA NA NA NA NA NA NA NA NA ...
 $ egg.diam        : num  NA NA NA NA NA NA NA NA NA NA ...
 $ lat             : num  65.7 65.7 65.7 65.7 65.7 ...
 $ long            : num  -64.8 -64.8 -64.8 -64.8 -64.8 ...
 $ depthA          : num  NA NA NA NA NA NA NA NA NA NA ...
 $ depthB          : num  NA NA NA NA NA NA NA NA NA NA ...
 $ depth           : num  NA NA NA NA NA NA NA NA NA NA ...
 $ net.len         : num  50 50 50 50 50 50 50 50 50 50 ...
 $ net.height      : num  1.83 1.83 1.83 1.83 1.83 1.83 1.83 1.83 1.83 1.83 ...
 $ wind.dir        : chr  NA NA NA NA ...
 $ wind.spd        : chr  "med" "med" "med" "med" ...
 $ temp.water      : num  NA NA NA NA NA NA NA NA NA NA ...
 $ temp.air        : num  NA NA NA NA NA NA NA NA NA NA ...
 $ sky             : chr  "overcast" "overcast" "overcast" "overcast" ...
 $ netset.time     : chr  "17:38:00" "17:38:00" "17:38:00" "17:38:00" ...
 $ netset.date     : POSIXct, format: "2012-03-03" "2012-03-03" "2012-03-03" ...
 $ netlift.time    : chr  "21:13:00" "21:13:00" "21:13:00" "21:13:00" ...
 $ netlift.date    : POSIXct, format: "2012-03-03" "2012-03-03" "2012-03-03" ...
 $ lake.zone       : chr  "benthic" "benthic" "benthic" "benthic" ...
```

```
 $ tag              : chr  NA NA NA NA ...
 $ remarks          : chr  NA "stomach parasites" NA NA ...
 $ fyear            : Factor w/ 8 levels "1983","1997",..: 6 6 6 6 6 6 6 6 6 6 ...


> dNU$FL
 [1] 421 510 638 399 705 646 476 520 579 507 632 178 324 365 411 657 539 433 654 701 588 560 550 540
[25] 500 430 630 500 530 500 620 530 600 530 560 550 550 540 470 650 600 670 640 560 610 570 510 550
[49] 620 550 600 650 570 560 590 630 500 550 610 600 590 670 500 600 610 600 620 500 560 500 540 520
 [ reached getOption("max.print") -- omitted 1973 entries ]
```