caterpillarcleaning.R

brittanycavazos

Fri Sep 30 15:20:45 2016

```
####Brittany Cavazos
####Caterpillar Data from Clay-Caterpillar Expt.
####Last edited:Sept. 29, 2016
####Goal:to "clean" data for easier analysis
#####################################################
#####################################################
library(readxl)
library(tidyr)
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
```

```
# read in caterpillar data

caterpillar<-read_excel("C:\\Users\\brittanycavazos\\Documents\\EEB_590\\ClayCaterpillarProject\\bcavaz
predation<-read_excel("C:\\Users\\brittanycavazos\\Documents\\EEB_590\\ClayCaterpillarProject\\bcavazos
# ^ had a problem w/ predation data with read_excel - had to add the col_types w/ 8 "text"s to have it


str(caterpillar)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1650 obs. of  9 variables:
##  $ island       : chr  "guam" "guam" "guam" "guam" ...
##  $ site         : chr  "anao" "anao" "anao" "anao" ...
##  $ date         : POSIXct, format: "2013-08-01" "2013-08-01" ...
##  $ type         : chr  "native" "native" "native" "native" ...
##  $ number       : chr  "1" "2" "3" "4" ...
##  $ result       : chr  "unpredated" "unpredated" "unpredated" "unpredated" ...
##  $ ground       : chr  "n" "n" "n" "n" ...
##  $ quality check: chr  "BRC 09/19/16" "BRC 09/19/16" "BRC 09/19/16" "BRC 09/19/16" ...
##  $ notes        : chr  NA NA NA NA ...
```

```
unique(caterpillar$result)
```

```
## [1] "unpredated" "predated"   "question"   "missing"
```

```r
# there are "missing (n=117)" and "question(n=11)" ones we may want to remove bc unsure if predated or
summary(as.factor(caterpillar$result))
```

```
##    missing   predated   question unpredated
##        118        277         13       1242
```

```r
colnames(caterpillar)[4]<-"habitat"

# it was being weird about subsetting so i'm doing it piecewise which i know is not the most efficient
caterpillar<-caterpillar[!caterpillar$result=="question",]
caterpillar<-caterpillar[!caterpillar$result=="missing",]

# we dont care about the analysis for Quality check and notes so we can take out those two columns
caterpillar<-caterpillar[,-9]
caterpillar<-caterpillar[,-8]

# so now that we took out the missing and question we can move on to fixing predation and merging them
names(caterpillar)
```

```
## [1] "island"  "site"    "date"    "habitat" "number"  "result"  "ground"
```

```r
names(predation)
```

```
## [1] "Island"  "Habitat" "Site"    "Number"  "Type"    "Notes"   ""
## [8] ""
```

```r
# get rid of two empty columns
predation<-predation[,1:6]
# first let's take out the ones that were mislabeled (these were ones that appeared to be assigned pred
predation<-predation[is.na(predation$Notes==T),]
# now we can take out the notes column because it's irrelevant
predation<-predation[,1:5]

names(predation)<-tolower(names(predation))
str(predation)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    289 obs. of  5 variables:
##  $ island : chr  "Guam" "Guam" "Guam" "Guam" ...
##  $ habitat: chr  "Native" "Native" "Native" "Native" ...
##  $ site   : chr  "Anaon" "Anaon" "Anaon" "Anaon" ...
##  $ number : chr  "68" "58" "18" "147" ...
##  $ type   : chr  "A" "A" "A" "B" ...
```

```r
predation$island<-as.factor(tolower(predation$island))
predation$habitat<-as.factor(tolower(predation$habitat))
predation$site<-as.factor(tolower(predation$site))

levels(predation$habitat)
```

```
## [1] "disturbed" "leucana"   "native"
```

```
# leucana = disturbed
predation$habitat <- gsub("leucana", "disturbed",predation$habitat)
# we also need to change the names around in site so they match each other
# in caterpillars - change anao to anao_n, ladtdn to ladt_n, marbo to marbo_d
# in predation - anaon to anao_n, blas to southblas, ladtn to ladt_n, marbod to marbo_d, tlp to twolove
caterpillar$site <- gsub("anao", "anao_n",caterpillar$site)
caterpillar$site <- gsub("ladtdn", "ladt_n",caterpillar$site)
caterpillar$site <- gsub("marbo", "marbo_d",caterpillar$site)

predation$site <- gsub("anaon", "anao_n",predation$site)
predation$site <- gsub("blas", "southblas",predation$site)
predation$site <- gsub("ladtn", "ladt_n",predation$site)
predation$site <- gsub("marbod", "marbo_d",predation$site)
predation$site <- gsub("tlp", "twolovers",predation$site)
predation$site <- gsub("tweks", "tweksberry",predation$site)

unique(caterpillar$site)
```

```
##  [1] "anao_n"     "calvo"      "grvp"       "ladtd"      "ladt_n"
##  [6] "marbo_d"    "naftan"     "palii"      "southblas"  "tweksberry"
## [11] "twolovers"
```

```
unique(predation$site)
```

```
##  [1] "anao_n"     "southblas"  "twolovers"  "marbo_d"    "ladt_n"
##  [6] "ladtd"      "naftan"     "palii"      "grvp"       "tweksberry"
## [11] "calvo"
```

```
#now merging them should work better -- before it was just assigning NAs to predation type even if even

summary(as.factor(predation$type))
```

```
##  * ??  A  B  C  D  E  F  G  H  I  J  L  M  N NP  O  P
## 39  9 12 28 90 19 16  7 13  7  7  2  1  3  1  6  8 21
```

```
# again, here there are some weird things - 41 occurances of a * (unrecognizable predation marking) and
# need to change the ones that =NP to switch result to unpredated

str(caterpillar)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1519 obs. of  7 variables:
##  $ island : chr  "guam" "guam" "guam" "guam" ...
##  $ site   : chr  "anao_n" "anao_n" "anao_n" "anao_n" ...
##  $ date   : POSIXct, format: "2013-08-01" "2013-08-01" ...
##  $ habitat: chr  "native" "native" "native" "native" ...
##  $ number : chr  "1" "2" "3" "4" ...
##  $ result : chr  "unpredated" "unpredated" "unpredated" "unpredated" ...
##  $ ground : chr  "n" "n" "n" "n" ...
```

```
caterpillar$island<-as.factor(caterpillar$island)
caterpillar$habitat<-as.factor(caterpillar$habitat)
caterpillar$site<-as.factor(caterpillar$site)

str(predation)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    289 obs. of  5 variables:
##  $ island : Factor w/ 3 levels "guam","rota",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ habitat: chr  "native" "native" "native" "native" ...
##  $ site   : chr  "anao_n" "anao_n" "anao_n" "anao_n" ...
##  $ number : chr  "68" "58" "18" "147" ...
##  $ type   : chr  "A" "A" "A" "B" ...
```

```
predation$island<-as.factor(predation$island)
predation$habitat<-as.factor(predation$habitat)
predation$site<-as.factor(predation$site)
predation$number<-as.character((predation$number))

predation$uniqueID<-paste(predation$site, predation$habitat, predation$number, sep = "-")
caterpillar$uniqueID<-paste(caterpillar$site, caterpillar$habitat, caterpillar$number, sep = "-")


#str of variables should match or it gives a warning message (predation habitat does not match caterpil
# so at this point, if we were to merge, we would want 1522 observations and 8 variables

caterpillardata<-left_join(caterpillar, predation, by = NULL)
```

```
## Joining, by = c("island", "site", "habitat", "number", "uniqueID")
```

```
# this worked

write.csv(caterpillardata, "C:\\Users\\brittanycavazos\\Documents\\EEB_590\\ClayCaterpillarProject\\bca
```