

Building Tidy Data

Ian Dinwoodie

2020-12-28

Loading the Data

Load the raw data and verify its dimensions and structure.

```
df <- read.csv('../data/dogs.csv', header=TRUE, skipNul=TRUE)
dim(df)
```

```
## [1] 1095  26
```

```
str(df)
```

```
## 'data.frame':  1095 obs. of  26 variables:
## $ Was...field.dog.name...acquired.at.12.weeks.or.younger.      : chr
## $ Is...field.dog.name...currently.at.least.1.year.old.         : chr
## $ How.many.years.old.is...field.dog.name...                   : int
## $ What.sex.is...field.dog.name...                               : chr
## $ Is...field.dog.name...spayed.or.neutered.                   : chr
## $ Did.you.take...field.dog.name...for.puppy.training.when.he.she.was.6.months.old.or.younger.: logi
## $ At.what.age.s...did.you.take...field.dog.name...for.training. : chr
## $ How.many.classes.did.you.and...field.dog.name...attend.      : chr
## $ At.puppy.training.classes..what.training.techniques.were.used. : chr
## $ What.restraining.training.devices.were.employed.            : chr
## $ Who.or.what.has...field.dog.name...acted.aggressively.toward. : chr
## $ What.sort.of.fears.and.or.anxiety.has...field.dog.name...had. : chr
## $ Who.does...field.dog.name...jumped.up.on.                   : chr
## $ When.has...field.dog.name...excessively.barked.             : chr
## $ What.type.of.feces.has...field.dog.name...eaten.            : chr
## $ What.sort.of.repetitive.behaviors.have.you.seen.with...field.dog.name... : chr
## $ When.has...field.dog.name...soiled.in.the.house.            : chr
## $ How.has...field.dog.name...soiled.in.the.house.             : chr
## $ Where.has...field.dog.name...soiled.in.the.house.           : chr
## $ In.what.repulsive.material.has...field.dog.name...rolled.    : chr
## $ In.what.ways.has...field.dog.name...been.overactive.hyperactive. : chr
## $ When.has...field.dog.name...been.destructive.               : chr
## $ Which.of.the.following.describes.how...field.dog.name...has.run.away.escaped. : chr
## $ Who.or.what.has...field.dog.name...mounted.humped.         : chr
## $ Do.you.have.another.dog.you.would.like.to.complete.the.questionnaire.for. : logi
## $ id                                                            : chr
```

We see that we have 1095 responses across 26 fields. The columns names are not quite serviceable in their current state, so we rename them for ease of use.

```
names <- c(
  'acq_12_wo_or_less',
  'at_least_1yo',
```

```

'age_yrs',
'sex',
'neutered',
'train_6mo_or_less',
'train_age',
'train_class_count',
'train_technique',
'restr_device',
'aggression',
'fear_anxiety',
'jumping',
'barking',
'coprophagia',
'compulsion',
'soil_when',
'soil_how',
'soil_where',
'rep_materials',
'hyperactive',
'destructive',
'escape',
'mounting',
'take_again',
'owner_id'
)
colnames(df) <- names
str(df)

```

```

## 'data.frame': 1095 obs. of 26 variables:
## $ acq_12_wo_or_less: chr "No" "Yes" "No" "Yes" ...
## $ at_least_1yo : chr "Yes" "Yes" "Yes" "Yes" ...
## $ age_yrs : int 7 9 10 5 NA 5 4 6 1 8 ...
## $ sex : chr "Male" "Female" "Female" "Male" ...
## $ neutered : chr "Yes" "Yes" "Yes" "Yes" ...
## $ train_6mo_or_less: logi FALSE FALSE FALSE TRUE FALSE FALSE ...
## $ train_age : chr "" "" "" "3 months or younger" ...
## $ train_class_count: chr "" "" "" "10+ classes" ...
## $ train_technique : chr "" "" "" "Rewarding techniques (e.g., treats, praise, pets)" ...
## $ restr_device : chr "" "" "" "Harness (around chest)" ...
## $ aggression : chr "Familiar people in the home, Unfamiliar dogs away from the home, Animals
## $ fear_anxiety : chr "Generalized anxiety, Fear of noises, Fear of thunderstorms, Fear of vete
## $ jumping : chr "" "" "" "Owners, Familiar people" ...
## $ barking : chr "" "" "" "" ...
## $ coprophagia : chr "" "Their own feces" "" "" ...
## $ compulsion : chr "when stressed climbs on top of furniture" "" "Sucking flank regions/blan
## $ soil_when : chr "" "" "" "" ...
## $ soil_how : chr "" "" "" "" ...
## $ soil_where : chr "" "" "" "" ...
## $ rep_materials : chr "" "" "Dead stuff" "" ...
## $ hyperactive : chr "" "" "" "" ...
## $ destructive : chr "" "" "Owner is away" "" ...
## $ escape : chr "" "" "Escaped when out, Returns home after escape" "" ...
## $ mounting : chr "" "" "" "" ...
## $ take_again : logi FALSE FALSE TRUE FALSE FALSE TRUE ...

```

```
## $ owner_id : chr "edd8d0889602b0c77117440a8defa33c" "75d0ac1234805817cb5659ac720c8fe3" "41"
```

Specifying Data Types

Continuous

We don't want to interpret every column as characters (chr), let's start by specifying the continuous variables.

```
df$age_yrs <- as.integer(df$age_yrs)
summary(df$age_yrs)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     NA's
##      0.00   4.00   7.00   7.18  10.00   65.00      70
```

Discrete

Factors

We convert a bulk of the columns to factors. Before the conversion, we need to deal with responses that are comma separated lists.

```
# For the training age we convert the responses from a list of ages take to the
# earliest attended age.
```

```
df$train_age <- ifelse(
  grepl('3 months or younger', df$train_age), '1-3 mo', df$train_age)
df$train_age <- ifelse(
  grepl('4 months', df$train_age), '4 mo', df$train_age)
df$train_age <- ifelse(
  grepl('5-6 months', df$train_age), '5-6 mo', df$train_age)
is.na(df$train_age) <- df$train_age == "I don't know"
df$train_age <- ordered(
  df$train_age, levels=c('1-3 mo', '4 mo', '5-6 mo'))
```

```
# Convert training technique to reward or punishment.
```

```
df$train_technique <- ifelse(
  grepl('Rewarding', df$train_technique), 'reward', df$train_technique)
df$train_technique <- ifelse(
  grepl('combination', df$train_technique), 'punish', df$train_technique)
df$train_technique <- ifelse(
  grepl('Tough love', df$train_technique), 'punish', df$train_technique)
df$train_technique <- ifelse(
  df$train_technique == 'reward' | df$train_technique == 'punish',
  df$train_technique, NA)
```

```
# Assign training class count to maximum selected option.
```

```
df$train_class_count <- ifelse(
  grepl('1-3', df$train_class_count), '1-3', df$train_class_count)
df$train_class_count <- ifelse(
  grepl('4-6', df$train_class_count), '4-6', df$train_class_count)
df$train_class_count <- ifelse(
  grepl('7-9', df$train_class_count), '7-9', df$train_class_count)
df$train_class_count <- ifelse(
  grepl('10+', df$train_class_count), '10+', df$train_class_count)
df$train_class_count <- ifelse(
  grepl('10+', df$train_class_count), '10+', df$train_class_count)
is.na(df$train_class_count) <- df$train_class_count == "I don't know"
```

```
df$train_class_count <- ordered(
  df$train_class_count, levels=c('1-3', '4-6', '7-9', '10+'))
```

Now we perform the conversion to factor data type.

```
factors <- c(
  'acq_12_wo_or_less',
  'at_least_1yo',
  'sex',
  'neutered',
  # 'train_age',
  # 'train_class_count',
  'train_technique',
  'restr_device',
  'aggression',
  'fear_anxiety',
  'jumping',
  'barking',
  'coprophagia',
  'compulsion',
  'soil_when',
  'soil_how',
  'soil_where',
  'rep_materials',
  'hyperactive',
  'destructive',
  'escape',
  'mounting',
  'owner_id'
)

for (c in factors) {
  df[, c] <- as.factor(df[, c])
}

str(df[, factors])
```

```
## 'data.frame': 1095 obs. of 21 variables:
## $ acq_12_wo_or_less: Factor w/ 3 levels "I don't know",...: 2 3 2 3 2 2 2 2 2 3 ...
## $ at_least_1yo : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 2 2 2 2 ...
## $ sex : Factor w/ 2 levels "Female","Male": 2 1 1 2 2 1 1 2 2 2 ...
## $ neutered : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
## $ train_technique : Factor w/ 2 levels "punish","reward": NA NA NA 2 NA NA NA NA NA ...
## $ restr_device : Factor w/ 64 levels "", "Buckle collar",...: 1 1 1 20 1 1 1 1 1 1 ...
## $ aggression : Factor w/ 226 levels "", "Animals other than dogs",...: 83 174 1 1 1 35 1 1 1 1
## $ fear_anxiety : Factor w/ 385 levels "", "afraid of little girls",...: 253 365 137 170 1 184 203
## $ jumping : Factor w/ 20 levels "", "Everyone",...: 1 1 1 10 1 1 1 1 1 1 ...
## $ barking : Factor w/ 108 levels "", "arriving home",...: 1 1 1 1 1 84 1 10 1 1 ...
## $ coprophagia : Factor w/ 34 levels "", "Cat feces",...: 1 29 1 1 1 10 1 1 29 1 ...
## $ compulsion : Factor w/ 126 levels "", "biting at foot",...: 125 1 98 1 1 1 1 1 78 1 ...
## $ soil_when : Factor w/ 53 levels "", "As a rescue he was not house trained in any way",...: 1
## $ soil_how : Factor w/ 9 levels "", "Both feces and urine",...: 1 1 1 1 1 2 1 1 2 1 ...
## $ soil_where : Factor w/ 11 levels "", "Anywhere",...: 1 1 1 1 1 2 1 1 8 1 ...
## $ rep_materials : Factor w/ 66 levels "", "Bird feces",...: 1 1 4 1 1 1 4 1 1 1 ...
```

```
## $ hyperactive      : Factor w/ 47 levels "", "At age 15, I no longer consider Abigail over active or
## $ destructive     : Factor w/ 11 levels "", "Confined in a small room",...: 1 1 4 1 1 4 1 1 1 1 ...
## $ escape           : Factor w/ 51 levels "", "1 time from house",...: 1 1 13 1 1 36 9 1 21 1 ...
## $ mounting         : Factor w/ 36 levels "", "\"Air humping\"",...: 1 1 1 1 1 4 1 1 1 1 ...
## $ owner_id         : Factor w/ 669 levels "0143addbe877065bb8d940e6e8901700",...: 624 311 185 185 51
```

Boolean

It's clear that some factor columns can be converted to boolean (i.e., logical).

```
df <- df %>%
  mutate(at_least_1yo = ifelse(at_least_1yo == 'Yes', TRUE, FALSE)) %>%
  mutate(neutered = ifelse(neutered == 'Yes', TRUE, FALSE))
```

Deriving Columns

We derive some columns for ease of use and improved clarity, especially when responses are comma separated lists.

```
df <- df %>%
  mutate(male = ifelse(sex == 'Male', FALSE, TRUE)) %>%
  mutate(device_used = ifelse(
    restr_device == "", NA, ifelse(
      grepl('No devices were employed', restr_device), FALSE, TRUE)))

# Derive a column for each restraining device.
df$buckle_collar <- ifelse(
  is.na(df$device_used), NA, ifelse(
    grepl('Buckle collar', df$restr_device), TRUE, FALSE))
df$martingale <- ifelse(
  is.na(df$device_used), NA, ifelse(
    grepl('Martingale collar', df$restr_device), TRUE, FALSE))
df$slip_collar <- ifelse(
  is.na(df$device_used), NA, ifelse(
    grepl('Nylon slip collar', df$restr_device), TRUE, FALSE))
df$shock_collar <- ifelse(
  is.na(df$device_used), NA, ifelse(
    grepl('Electric shock collar', df$restr_device), TRUE, FALSE))
df$harness <- ifelse(
  is.na(df$device_used), NA, ifelse(
    grepl('Harness', df$restr_device), TRUE, FALSE))
df$harness <- ifelse(
  is.na(df$device_used), NA, ifelse(
    grepl('harness', df$restr_device), TRUE, df$harness))
df$head_halter <- ifelse(
  is.na(df$device_used), NA, ifelse(
    grepl('Head halter', df$restr_device), TRUE, FALSE))
df$choke_collar <- ifelse(
  is.na(df$device_used), NA, ifelse(
    grepl('Metal \"choke\" collar', df$restr_device), TRUE, FALSE))
df$prong_collar <- ifelse(
  is.na(df$device_used), NA, ifelse(
    grepl('Prong collar', df$restr_device), TRUE, FALSE))
df$no_devices <- ifelse(
```

```
is.na(df$device_used), NA, ifelse(
  grepl('No devices were employed', df$restr_device), TRUE, FALSE))
```

Response Complexity Reductions

To start, we reduce the behavior problems to boolean indicators.

```
df <- df %>%
  mutate(aggression = ifelse(aggression == "", FALSE, TRUE)) %>%
  mutate(fear_anxiety = ifelse(fear_anxiety == "", FALSE, TRUE)) %>%
  mutate(jumping = ifelse(jumping == "", FALSE, TRUE)) %>%
  mutate(barking = ifelse(barking == "", FALSE, TRUE)) %>%
  mutate(coprophagia = ifelse(coprophagia == "", FALSE, TRUE)) %>%
  mutate(compulsion = ifelse(compulsion == "", FALSE, TRUE)) %>%
  mutate(house_soiling = ifelse(
    soil_when != "" | soil_how != "" | soil_where != "", FALSE, TRUE)) %>%
  mutate(rep_materials = ifelse(rep_materials == "", FALSE, TRUE)) %>%
  mutate(hyperactive = ifelse(hyperactive == "", FALSE, TRUE)) %>%
  mutate(destructive = ifelse(destructive == "", FALSE, TRUE)) %>%
  mutate(escape = ifelse(escape == "", FALSE, TRUE)) %>%
  mutate(mounting = ifelse(mounting == "", FALSE, TRUE))

str(df)
```

```
## 'data.frame': 1095 obs. of 38 variables:
## $ acq_12_wo_or_less: Factor w/ 3 levels "I don't know",...: 2 3 2 3 2 2 2 2 3 ...
## $ at_least_1yo : logi TRUE TRUE TRUE TRUE FALSE TRUE ...
## $ age_yrs : int 7 9 10 5 NA 5 4 6 1 8 ...
## $ sex : Factor w/ 2 levels "Female","Male": 2 1 1 2 2 1 1 2 2 2 ...
## $ neutered : logi TRUE TRUE TRUE TRUE TRUE TRUE ...
## $ train_6mo_or_less: logi FALSE FALSE FALSE TRUE FALSE FALSE ...
## $ train_age : Ord.factor w/ 3 levels "1-3 mo"<"4 mo"<...: NA NA NA 1 NA NA NA NA NA ...
## $ train_class_count: Ord.factor w/ 4 levels "1-3"<"4-6"<"7-9"<...: NA NA NA 4 NA NA NA NA NA ...
## $ train_technique : Factor w/ 2 levels "punish","reward": NA NA NA 2 NA NA NA NA NA ...
## $ restr_device : Factor w/ 64 levels "", "Buckle collar",...: 1 1 1 20 1 1 1 1 1 1 ...
## $ aggression : logi TRUE TRUE FALSE FALSE FALSE TRUE ...
## $ fear_anxiety : logi TRUE TRUE TRUE TRUE FALSE TRUE ...
## $ jumping : logi FALSE FALSE FALSE TRUE FALSE FALSE ...
## $ barking : logi FALSE FALSE FALSE FALSE FALSE TRUE ...
## $ coprophagia : logi FALSE TRUE FALSE FALSE FALSE TRUE ...
## $ compulsion : logi TRUE FALSE TRUE FALSE FALSE FALSE ...
## $ soil_when : Factor w/ 53 levels "", "As a rescue he was not house trained in any way",...: 1 ...
## $ soil_how : Factor w/ 9 levels "", "Both feces and urine",...: 1 1 1 1 1 2 1 1 2 1 ...
## $ soil_where : Factor w/ 11 levels "", "Anywhere",...: 1 1 1 1 1 2 1 1 8 1 ...
## $ rep_materials : logi FALSE FALSE TRUE FALSE FALSE FALSE ...
## $ hyperactive : logi FALSE FALSE FALSE FALSE FALSE TRUE ...
## $ destructive : logi FALSE FALSE TRUE FALSE FALSE TRUE ...
## $ escape : logi FALSE FALSE TRUE FALSE FALSE TRUE ...
## $ mounting : logi FALSE FALSE FALSE FALSE FALSE TRUE ...
## $ take_again : logi FALSE FALSE TRUE FALSE FALSE TRUE ...
## $ owner_id : Factor w/ 669 levels "0143addbe877065bb8d940e6e8901700",...: 624 311 185 185 51 ...
## $ male : logi FALSE TRUE TRUE FALSE FALSE TRUE ...
## $ device_used : logi NA NA NA TRUE NA NA ...
```

```
## $ buckle_collar      : logi  NA NA NA FALSE NA NA ...
## $ martingale         : logi  NA NA NA FALSE NA NA ...
## $ slip_collar       : logi  NA NA NA FALSE NA NA ...
## $ shock_collar      : logi  NA NA NA FALSE NA NA ...
## $ harness           : logi  NA NA NA TRUE NA NA ...
## $ head_halter       : logi  NA NA NA FALSE NA NA ...
## $ choke_collar      : logi  NA NA NA FALSE NA NA ...
## $ prong_collar      : logi  NA NA NA FALSE NA NA ...
## $ no_devices        : logi  NA NA NA FALSE NA NA ...
## $ house_soiling     : logi   TRUE TRUE TRUE TRUE TRUE FALSE ...
```

Dropping Excess Data

Applying Inclusion Criteria

```
df <- filter(df, at_least_1yo == TRUE, age_yrs >= 1, age_yrs <= 35)
dim(df)
```

```
## [1] 1023   38
```

Dropping Columns

Drop columns that serve no purpose with the analysis.

```
df <- subset(df, select=-c(
  take_again, # survey software logic variable
  soil_when,
  soil_how,
  soil_where,
  at_least_1yo, # survey software logic variable
  sex, # replaced with a male column
  restr_device, # devices moved into their own columns
  no_devices # mirrors the device_used column
))
```

Final Summary

Take a last look at the data before saving it to disk.

```
dim(df)
```

```
## [1] 1023   30
```

```
summary(df)
```

```
##      acq_12_wo_or_less      age_yrs      neutered      train_6mo_or_less
## I don't know: 17      Min.   : 1.000      Mode :logical      Mode :logical
## No           :449      1st Qu.: 4.000      FALSE:132      FALSE:529
## Yes          :557      Median : 7.000      TRUE :891       TRUE :494
##                                     Mean   : 7.131
##                                     3rd Qu.:10.000
##                                     Max.   :19.000
##
##      train_age      train_class_count      train_technique      aggression      fear_anxiety
## 1-3 mo:234      1-3 : 49                punish: 54                Mode :logical      Mode :logical
```

```

## 4 mo :130 4-6 :120 reward:440 FALSE:474 FALSE:310
## 5-6 mo:118 7-9 : 72 NA's :529 TRUE :549 TRUE :713
## NA's :541 10+ :242
## NA's:540
##
##
## jumping barking coprophagia compulsion
## Mode :logical Mode :logical Mode :logical Mode :logical
## FALSE:793 FALSE:806 FALSE:642 FALSE:769
## TRUE :230 TRUE :217 TRUE :381 TRUE :254
##
##
##
## rep_materials hyperactive destructive escape
## Mode :logical Mode :logical Mode :logical Mode :logical
## FALSE:595 FALSE:907 FALSE:892 FALSE:793
## TRUE :428 TRUE :116 TRUE :131 TRUE :230
##
##
##
## mounting owner_id male
## Mode :logical 3ea182741999dd54cb902c478ba2704c: 8 Mode :logical
## FALSE:833 1b9b35f5434de88ff7f3ff4b0e371d48: 7 FALSE:526
## TRUE :190 796cf2f6f66cf06329ecc6067d7419f0: 6 TRUE :497
## a5069b3d48cbac2d77080428c7d8d315: 6
## f9968086714b82f1c1c87019d1187507: 6
## 0d29a6dde9e38788ba6a480bf902fb53: 4
## (Other) :986
## device_used buckle_collar martingale slip_collar
## Mode :logical Mode :logical Mode :logical Mode :logical
## FALSE:62 FALSE:259 FALSE:404 FALSE:449
## TRUE :432 TRUE :235 TRUE :90 TRUE :45
## NA's :529 NA's :529 NA's :529 NA's :529
##
##
##
## shock_collar harness head_halter choke_collar
## Mode :logical Mode :logical Mode :logical Mode :logical
## FALSE:485 FALSE:345 FALSE:468 FALSE:467
## TRUE :9 TRUE :149 TRUE :26 TRUE :27
## NA's :529 NA's :529 NA's :529 NA's :529
##
##
##
## prong_collar house_soiling
## Mode :logical Mode :logical
## FALSE:461 FALSE:225
## TRUE :33 TRUE :798
## NA's :529
##
##
##

```


Saving the Tidy Data

Save the data to a file in RDS format so that the data types are saved and so that the output is compressed.

```
saveRDS(df, '../data/tidy.Rds')
```