Exploratory Data Analysis

Ian Dinwoodie

2021-01-24

Loading the Data

Load the raw data and verify its dimensions and structure.

```
df <- readRDS('../data/tidy.Rds')
dim(df)
```

```
## [1] 1023   35
```

```
str(df)
```

```
## 'data.frame':    1023 obs. of  35 variables:
##  $ acq_12_wo_or_less  : logi  FALSE TRUE FALSE TRUE FALSE FALSE ...
##  $ age_yrs            : int  7 9 10 5 5 4 6 1 8 11 ...
##  $ neutered           : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ train_6mo_or_less  : logi  FALSE FALSE FALSE TRUE FALSE FALSE ...
##  $ train_class_count  : Ord.factor w/ 4 levels "1-3"<"4-6"<"7-9"<..: NA NA NA 4 NA NA NA NA NA 3 ...
##  $ train_technique    : Factor w/ 2 levels "punish","reward": NA NA NA 2 NA NA NA NA NA 2 ...
##  $ aggression         : logi  TRUE TRUE FALSE FALSE TRUE FALSE ...
##  $ fear_anxiety       : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ jumping            : logi  FALSE FALSE FALSE TRUE FALSE FALSE ...
##  $ barking            : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ coprophagia        : logi  FALSE TRUE FALSE FALSE TRUE FALSE ...
##  $ compulsion         : logi  TRUE FALSE TRUE FALSE FALSE FALSE ...
##  $ rep_materials      : logi  FALSE FALSE TRUE FALSE FALSE TRUE ...
##  $ hyperactive        : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ destructive        : logi  FALSE FALSE TRUE FALSE TRUE FALSE ...
##  $ escape             : logi  FALSE FALSE TRUE FALSE TRUE TRUE ...
##  $ mounting           : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ owner_id           : Factor w/ 669 levels "0143addbe877065bb8d940e6e8901700",..: 624 311 185 185
##  $ train_1_3_mo       : Factor w/ 2 levels "FALSE","TRUE": NA NA NA 2 NA NA NA NA NA 1 ...
##  $ train_4_mo         : Factor w/ 2 levels "FALSE","TRUE": NA NA NA 1 NA NA NA NA NA 2 ...
##  $ train_5_6_mo       : Factor w/ 2 levels "FALSE","TRUE": NA NA NA 1 NA NA NA NA NA 1 ...
##  $ train_start_age    : Ord.factor w/ 3 levels "1-3 mo"<"4 mo"<..: NA NA NA 1 NA NA NA NA NA 2 ...
##  $ male               : logi  FALSE TRUE TRUE FALSE TRUE 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 ...
##  $ house_soiling      : logi  TRUE TRUE TRUE TRUE FALSE TRUE ...
##  $ adj_train_technique: Factor w/ 2 levels "punish","reward": NA NA NA 2 NA NA NA NA NA 2 ...
##  $ punish_device      : Factor w/ 2 levels "FALSE","TRUE": NA NA NA 1 NA NA NA NA NA 1 ...
```

Basic Exploration

Owner Identifier

The owner identifier is used to calculate the number of owners.

```
# Number of unique owners after inclusion criteria.
length(unique(df$owner_id))
```

```
## [1] 641
```

It is also used to calculate the number of dogs per household.

```
summary(plyr::count(df, 'owner_id'))
```

```
##                               owner_id        freq
##  0180dd62878f2d494db4e6aae4695386:  1   Min.   :1.000
##  018b0b08b0a8dbc63f58c47b0c94d2e4:  1   1st Qu.:1.000
##  01bbe34d450b00b4fc3ce4b319986b81:  1   Median :1.000
##  01f09881fb038ee28ab0ef02aa80d87a:  1   Mean   :1.596
##  01f5bf70d07b1e05f2c5ba2fdb6c40fb:  1   3rd Qu.:2.000
##  02c983a2def62e515889b3f6657b212c:  1   Max.   :8.000
##  (Other)                         :635
```

We see the median number of dogs per household is 1 (range: 1 to 8). Now we can drop the column to simply the data set.

```
df <- subset(df, select=-c(owner_id))
```

Overview of Data Set

Before we look at the data, let's add a basic behavior problem indicator column.

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

summary(df$behav_problem)
```

```
##    Mode   FALSE    TRUE
## logical       7    1016
```

Let's take a look at the data set as we head toward analysis.

```
summary(df)
```

```
##  acq_12_wo_or_less    age_yrs        neutered       train_6mo_or_less
##  Mode :logical     Min.   : 1.000   Mode :logical   Mode :logical
##  FALSE:449         1st Qu.: 4.000   FALSE:132       FALSE:529
##  TRUE :557         Median : 7.000   TRUE :891       TRUE :494
##  NA's :17          Mean   : 7.131
##                    3rd Qu.:10.000
##                    Max.   :19.000
##  train_class_count train_technique aggression      fear_anxiety
##  1-3 : 49          punish: 54      Mode :logical   Mode :logical
##  4-6 :120          reward:440      FALSE:474       FALSE:310
##  7-9 : 72          NA's  :529      TRUE :549       TRUE :713
##  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       train_1_3_mo train_4_mo  train_5_6_mo train_start_age
##  Mode :logical   FALSE:248    FALSE:267   FALSE:256    1-3 mo:234
##  FALSE:833       TRUE :234    TRUE :215   TRUE :226    4 mo  :130
##  TRUE :190       NA's :541    NA's :541   NA's :541    5-6 mo:118
##                                                        NA's  :541
##
##
##     male         device_used     buckle_collar   martingale
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:526       FALSE:62        FALSE:259       FALSE:404
##  TRUE :497       TRUE :432       TRUE :235       TRUE :90
##                  NA's :529       NA's :529       NA's :529
##
##
##  slip_collar     shock_collar     harness        head_halter
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:449       FALSE:485       FALSE:345       FALSE:468
##  TRUE :45        TRUE :9         TRUE :149       TRUE :26
##  NA's :529       NA's :529       NA's :529       NA's :529
##
##
##  choke_collar    prong_collar    house_soiling   adj_train_technique
##  Mode :logical   Mode :logical   Mode :logical   punish:178
##  FALSE:467       FALSE:461       FALSE:225       reward:316
##  TRUE :27        TRUE :33        TRUE :798       NA's  :529
##  NA's :529       NA's :529
##
##
##  punish_device behav_problem
##  FALSE:316     Mode :logical
##  TRUE :178     FALSE:7
##  NA's :529     TRUE :1016
##
```

```
##
##
```

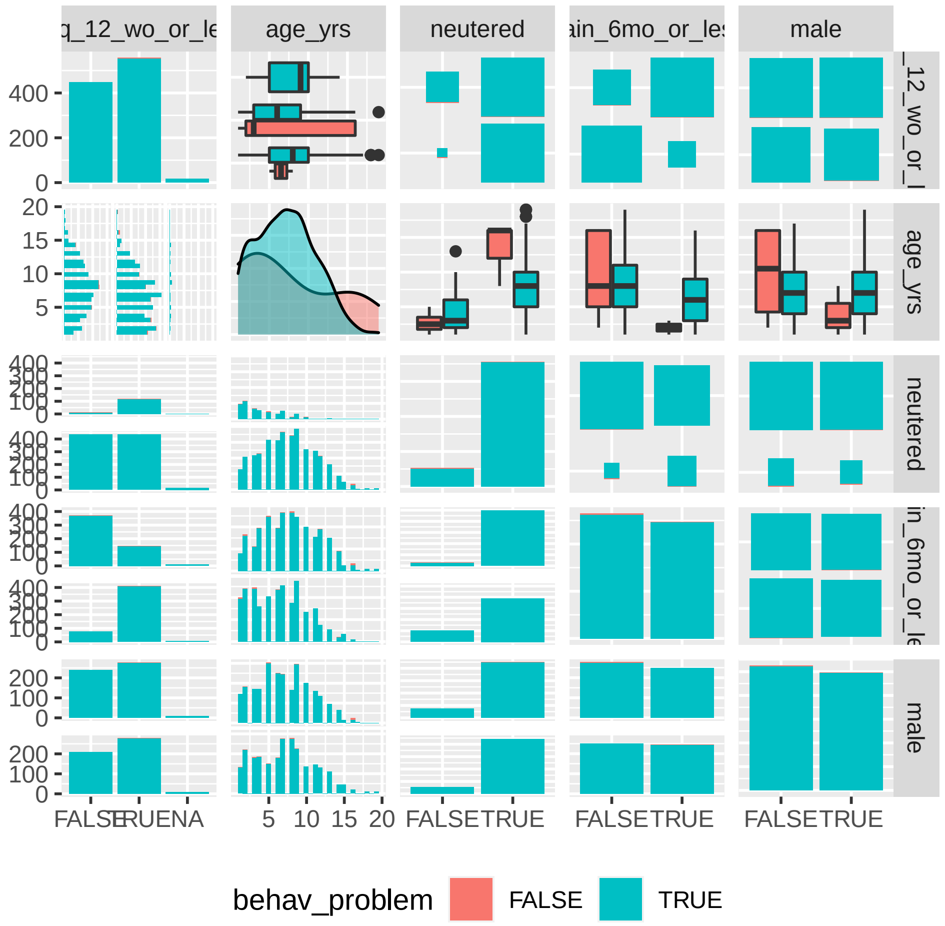
Notable observations:

- Median dog age is 7 yrs (range: 1 to 19 yrs).
  - More than half (54.4%) were acquired at 12 weeks or less.
- A majority (87.1%) of dogs were neutered.
- The gender split is nearly even with 48.6% males.
- About half of the dogs (48.3%) attended training at 6 months old or earlier (i.e., puppy training).
  - About half (47.4%) of which started attending in the 1-3 month range.
  - A majority (87.4%) of the dogs that attended puppy training were subject to some form of restraining device.
    - The buckle collar was the most popular device at 47.6% usage.
    - The shock collar was the least popular at 1.8% usage.
  - A vast majority (89.1%) were believed to have been subjected to reward-based training.
    - Correcting for punishing restraint devices, only 64.0% were truly subject to reward based training; a 25.1% difference!
- A vast majority of dogs (99.3%) were reported to exhibit at least one type of problematic behavior.
  - The top 3 most frequent behavior problems were house soiling, fear/anxiety, aggression.
  - The 3 least frequent behavior problems were hyperactivity, destruction, and mounting.

We create pairwise scatter plots for columns that all participants were presented (i.e., no NA responses) and we exclude the individual behavior problem columns for brevity.

```r
df %>%
  ggpairs(columns=c('acq_12_wo_or_less', 'age_yrs', 'neutered',
                    'train_6mo_or_less', 'male'),
          mapping=ggplot2::aes(color=behav_problem),
          diag=list(discrete='barDiag',
                    continuous=wrap('densityDiag', alpha=0.5)),
          legend=1,
          progress=FALSE) +
  theme(legend.position='bottom')
```

```
## Warning: Removed 17 rows containing non-finite values (stat_g_gally_count).

## Warning: Removed 17 rows containing non-finite values (stat_g_gally_count).

## Warning: Removed 17 rows containing non-finite values (stat_g_gally_count).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

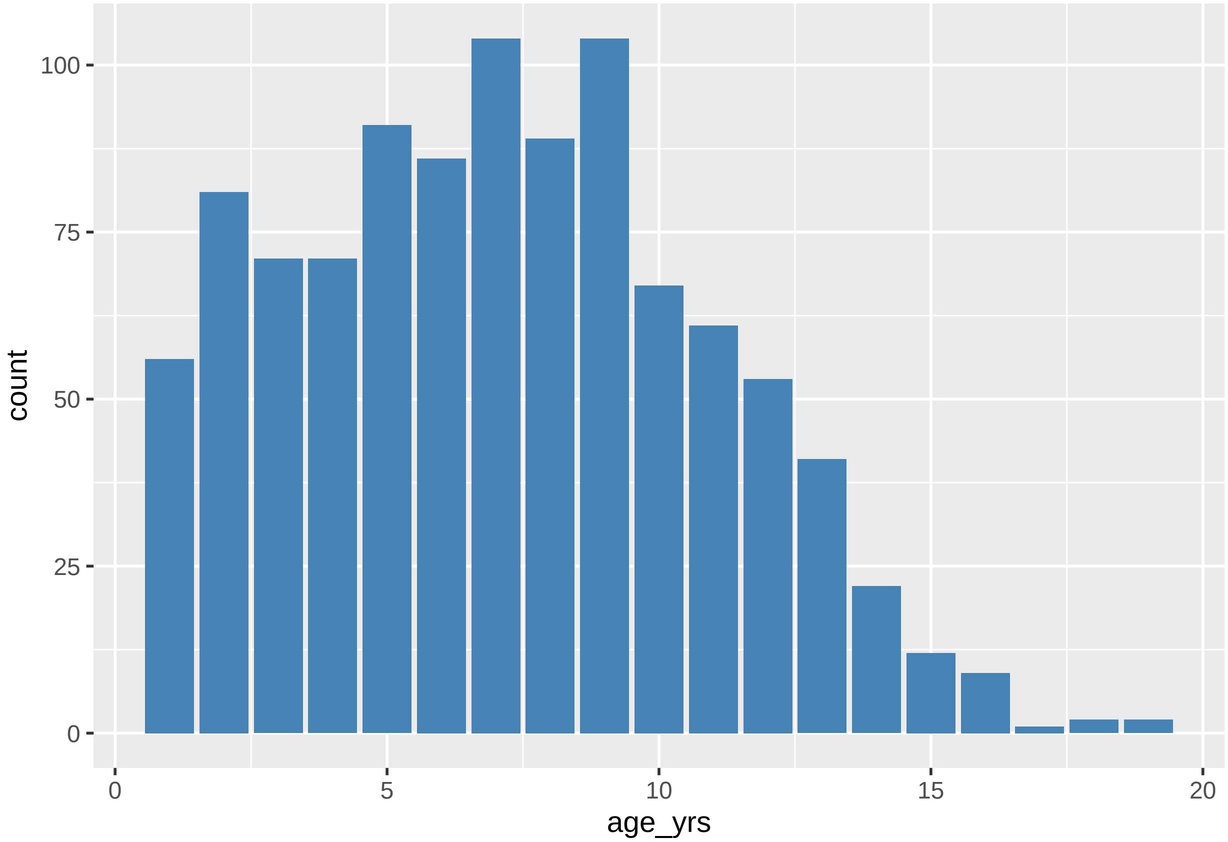


Our control group consists of the dogs that did not attend puppy training. We can compare variable distributions across the experimental and control groups by looking at the graphs along the `train_6mo_or_less` row. Thankfully, we see that the distributions between the two groups for the plotted columns are roughly equivalent.

Continuous Variables

The age of the dog is the only continuous variable we are working with.

```r
ggplot(df, aes(age_yrs)) + geom_bar(fill='steelblue')
```

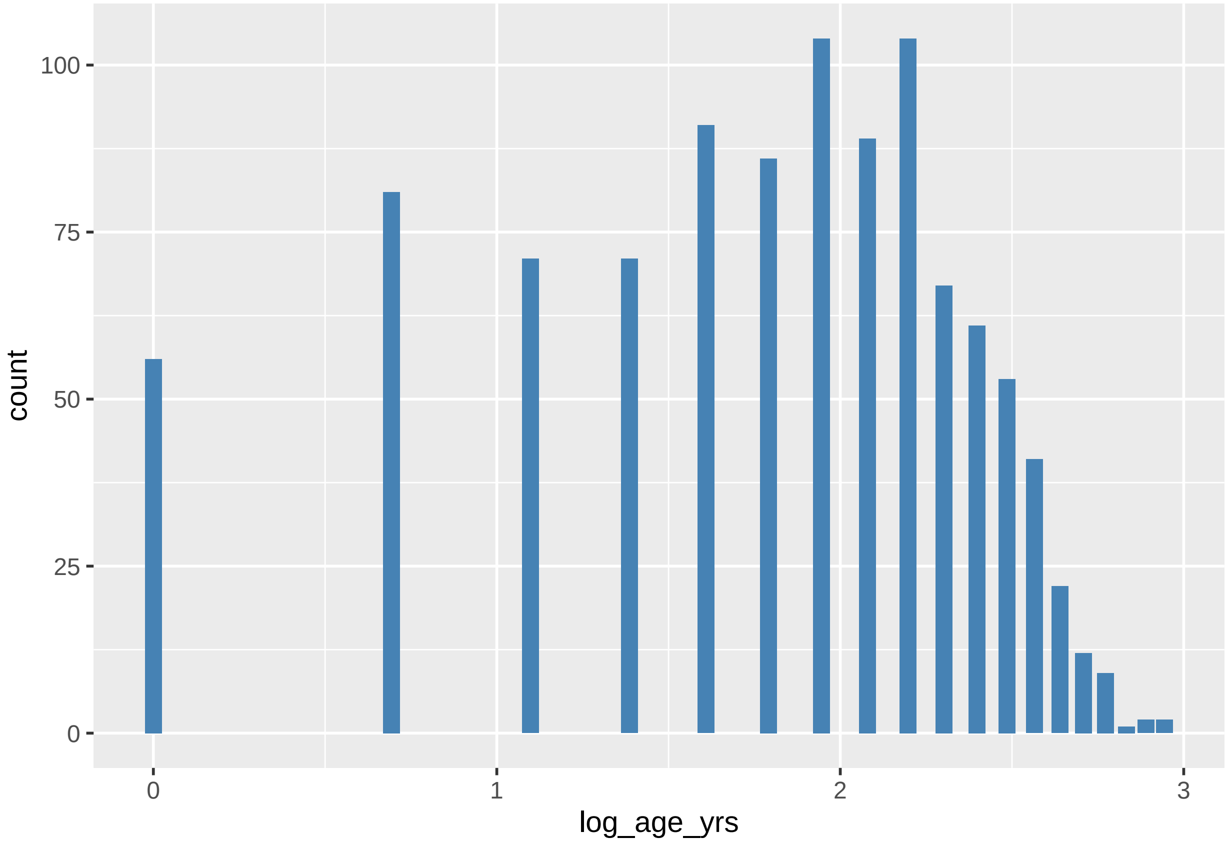


```
skewness(df$age_yrs)
```

```
## [1] 0.2669543
```

We see a slight right skew in the plot. Let's try to center it by applying a log transform.

```
df <- df %>%
  mutate(log_age_yrs = log(age_yrs))
ggplot(df, aes(log_age_yrs)) + geom_bar(fill='steelblue')
```



```r
skewness(df$log_age_yrs)
```

```
## [1] -0.9868858
```

```r
df <- subset(df, select=-c(log_age_yrs))
```

We see that the log transform resulted in a greater absolute skew, so we drop the transformed column and rely on the original.

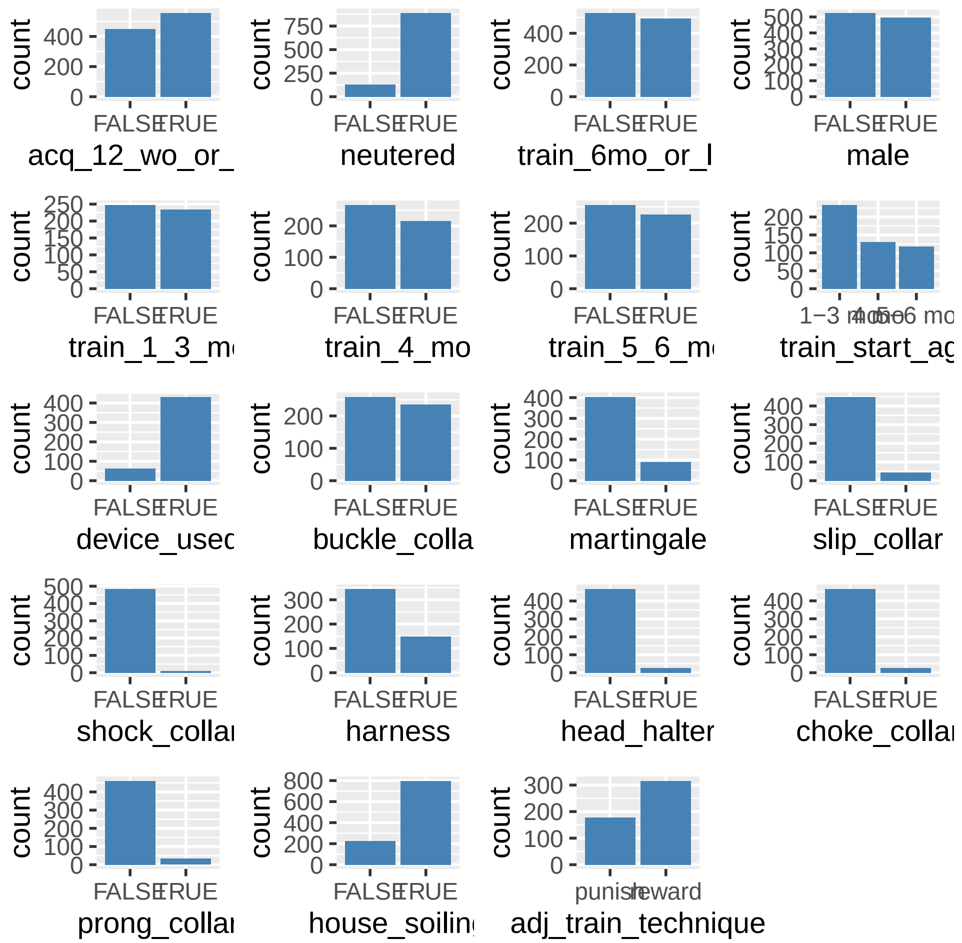
Discrete Variables

Independent Variables

```r
vars <- c(
  'acq_12_wo_or_less',
  'neutered',
  'train_6mo_or_less',
  'male',
  'train_1_3_mo',
  'train_4_mo',
  'train_5_6_mo',
  'train_start_age',
  'device_used',
  'buckle_collar',
  'martingale',
  'slip_collar',
  'shock_collar',
  'harness',
  'head_halter',
  'choke_collar',
  'prong_collar',
  'house_soiling',
  'adj_train_technique'
)

plot_list <- list()
for (i in 1:length(vars)) {
  col <- vars[i]
  p <- df %>%
    select(col) %>%
    drop_na(col) %>%
    ggplot(aes_string(x = col)) +
    geom_bar(fill='steelblue')
  plot_list[[i]] <- p
}
```

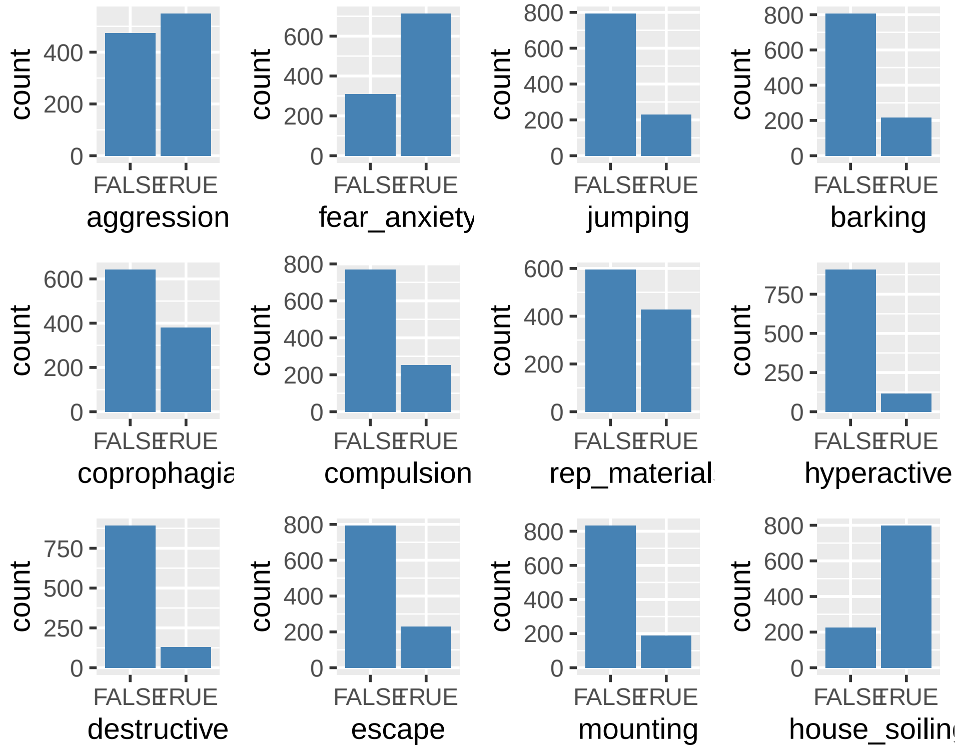
```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(col)` instead of `col` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

```r
ggarrange(plotlist=plot_list, ncol=4, nrow=5)
```



Dependent Variables

```
outcomes <- c(
  'aggression',
  'fear_anxiety',
  'jumping',
  'barking',
  'coprophagia',
  'compulsion',
  'rep_materials',
  'hyperactive',
  'destructive',
  'escape',
  'mounting',
  'house_soiling'
)

plot_list <- list()
for (i in 1:length(outcomes)) {
  col <- outcomes[i]
  p <- df %>%
    select(col) %>%
    drop_na(col) %>%
    ggplot(aes_string(x = col)) +
    geom_bar(fill='steelblue')
  plot_list[[i]] <- p
}
ggarrange(plotlist=plot_list, ncol=4, nrow=3)
```



Exploring Trends and Relationships

Sex and Neuter Status

It's common to want to know the split of neuter status by sex, so let's generate those numbers now.

```
xtab <- xtabs(~male+neutered, data=df)
print(xtab)
```

```
##        neutered
## male    FALSE TRUE
##   FALSE    76  450
##   TRUE     56  441
```

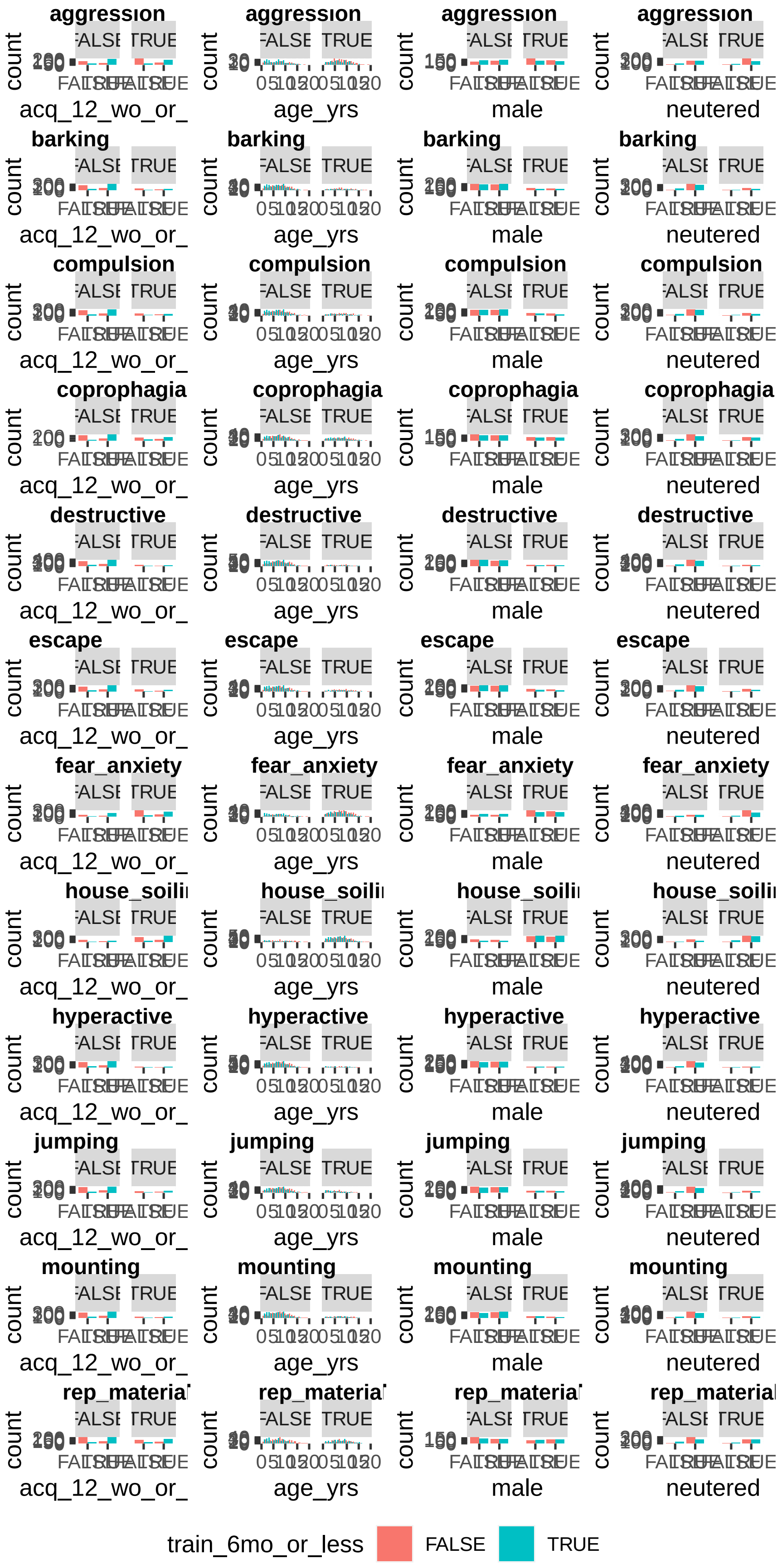
Control vs Experimental Group

Our control group consists of the dogs who did not attend puppy training and our experimental group consists of those who did. Let's look at the variables common to both groups with the plot color indicating the presence of a behavior problem. Since we know a vast majority of dogs have at least one behavior problem, we need to look for trends in individual behavior problems for the plots to be useful.

```
# Generate plots for each attribute split by a simple predictor.
pred <- 'train_6mo_or_less'
attribs <- c(
  'acq_12_wo_or_less',
  'age_yrs',
  'male',
  'neutered'
)
attribs <- sort(attribs)
outcomes <- sort(outcomes)

plot_list <- list()
cnt <- 1
labels <- NULL
for (i in 1:length(outcomes)) {
  outcome <- outcomes[i]
  for (j in 1:length(attribs)) {
    attrib <- attribs[j]
    p <- df %>%
      drop_na(attrib) %>%
      select(attrib, outcome, pred) %>%
      ggplot(aes_string(x=attrib, fill=pred)) +
      geom_bar(position = position_dodge(0.9)) +
      labs(fill=pred) +
      theme(legend.position='none') +
      facet_grid(as.formula(paste0('.~', outcome)))
    plot_list[[cnt]] <- p
    cnt <- cnt + 1
    labels <- c(labels, outcome)
  }
}
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(attrib)` instead of `attrib` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.

## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(outcome)` instead of `outcome` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.

## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(pred)` instead of `pred` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

```
ggarrange(plotlist=plot_list, ncol=4, nrow=12, common.legend=TRUE,
          font.label=list(size=10), vjust=0.75, legend='bottom', labels=labels)
```


aggression
barking
compulsion
coprophagia
destructive
escape
fear_anxiety
house_soili
hyperactive
jumping
mounting
rep_materia
rep_material
FALSE
TRUE
count
acq_12_wo_or_
age_yrs
male
neutered
train_6mo_or_less
FALSE
TRUE

Note: For each single plot the behavior problem is indicated by the label in the top left corner. The left facet is the group of dogs without the behavior problem and the right are the dogs with the behavior problem. Within each facet the color indicates control (red) or experiment (blue) grouping.

Within the Experimental Group

Within the experimental group we are curious to see the impact of various training techniques and restrain devices on behavior problem occurrence. We start by isolating the experimental group.

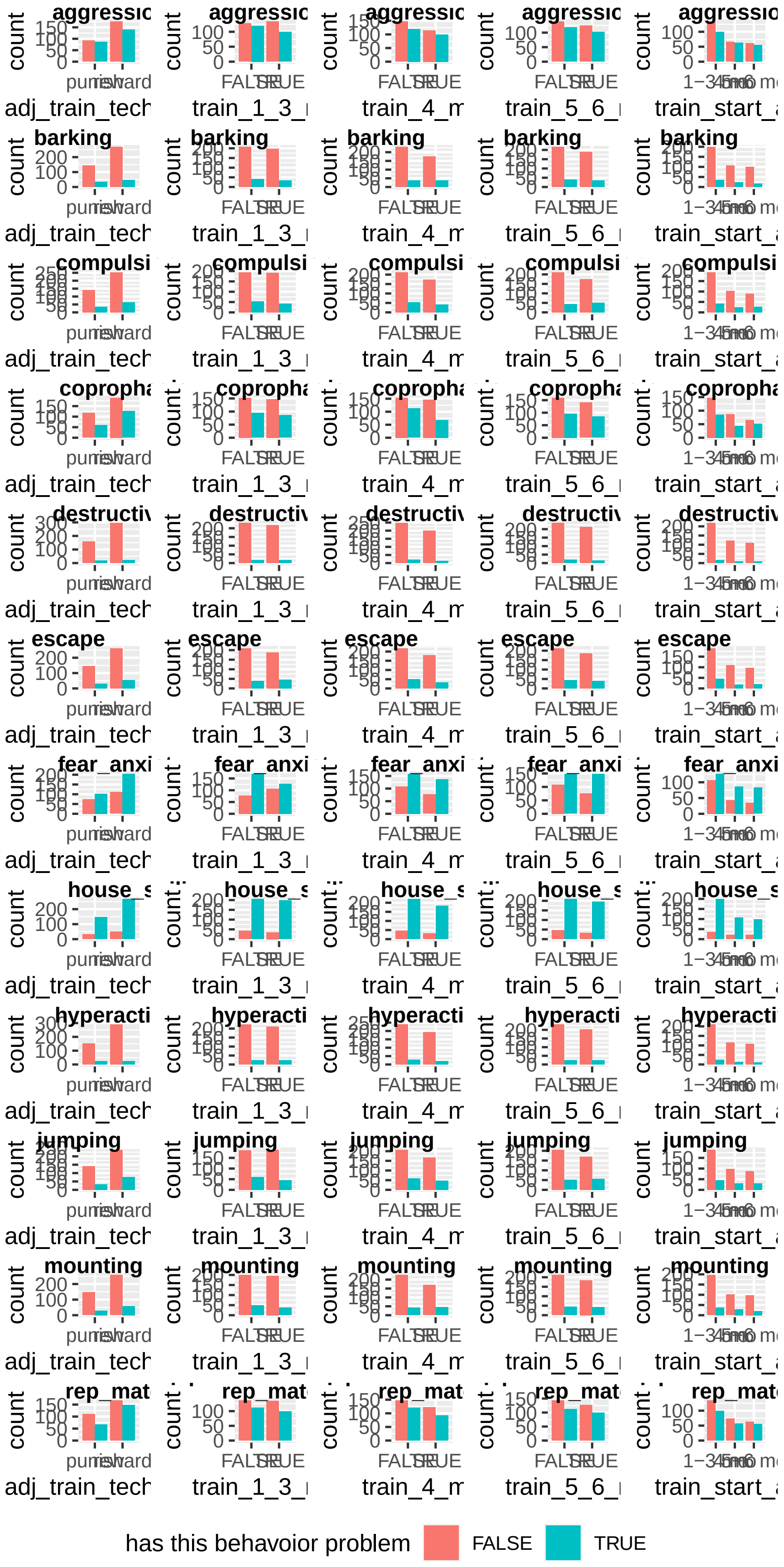
```
df_exp <- df %>%
  filter(train_6mo_or_less == TRUE)
summary(df_exp)
```

```
##  acq_12_wo_or_less    age_yrs        neutered       train_6mo_or_less
##  Mode :logical     Min.   : 1.000   Mode :logical   Mode:logical
##  FALSE:78          1st Qu.: 3.000   FALSE:103       TRUE:494
##  TRUE :410         Median : 6.000   TRUE :391
##  NA's :6           Mean   : 6.368
##                    3rd Qu.: 9.000
##                    Max.   :16.000
##  train_class_count train_technique aggression      fear_anxiety
##  1-3 : 49          punish: 54      Mode :logical   Mode :logical
##  4-6 :120          reward:440      FALSE:267       FALSE:186
##  7-9 : 72                          TRUE :227       TRUE :308
##  10+ :242
##  NA's: 11
##
##   jumping         barking        coprophagia     compulsion
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:382       FALSE:412       FALSE:308       FALSE:394
##  TRUE :112       TRUE :82        TRUE :186       TRUE :100
##
##
##
##  rep_materials   hyperactive     destructive       escape
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:278       FALSE:442       FALSE:455       FALSE:407
##  TRUE :216       TRUE :52        TRUE :39        TRUE :87
##
##
##
##   mounting       train_1_3_mo train_4_mo  train_5_6_mo train_start_age
##  Mode :logical   FALSE:248    FALSE:267   FALSE:256    1-3 mo:234
##  FALSE:405       TRUE :234    TRUE :215   TRUE :226    4 mo  :130
##  TRUE :89        NA's : 12    NA's : 12   NA's : 12    5-6 mo:118
##                                                        NA's  : 12
##
##
##     male          device_used     buckle_collar   martingale
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:248       FALSE:62        FALSE:259       FALSE:404
##  TRUE :246       TRUE :432       TRUE :235       TRUE :90
##
##
```

```
##
##  slip_collar     shock_collar     harness        head_halter
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:449       FALSE:485       FALSE:345       FALSE:468
##  TRUE :45        TRUE :9         TRUE :149       TRUE :26
##
##
##
##  choke_collar    prong_collar    house_soiling   adj_train_technique
##  Mode :logical   Mode :logical   Mode :logical   punish:178
##  FALSE:467       FALSE:461       FALSE:81        reward:316
##  TRUE :27        TRUE :33        TRUE :413
##
##
##
##  punish_device behav_problem
##  FALSE:316     Mode :logical
##  TRUE :178     FALSE:2
##                TRUE :492
##
##
##
```

Now we look at the impact of training age and frequency.

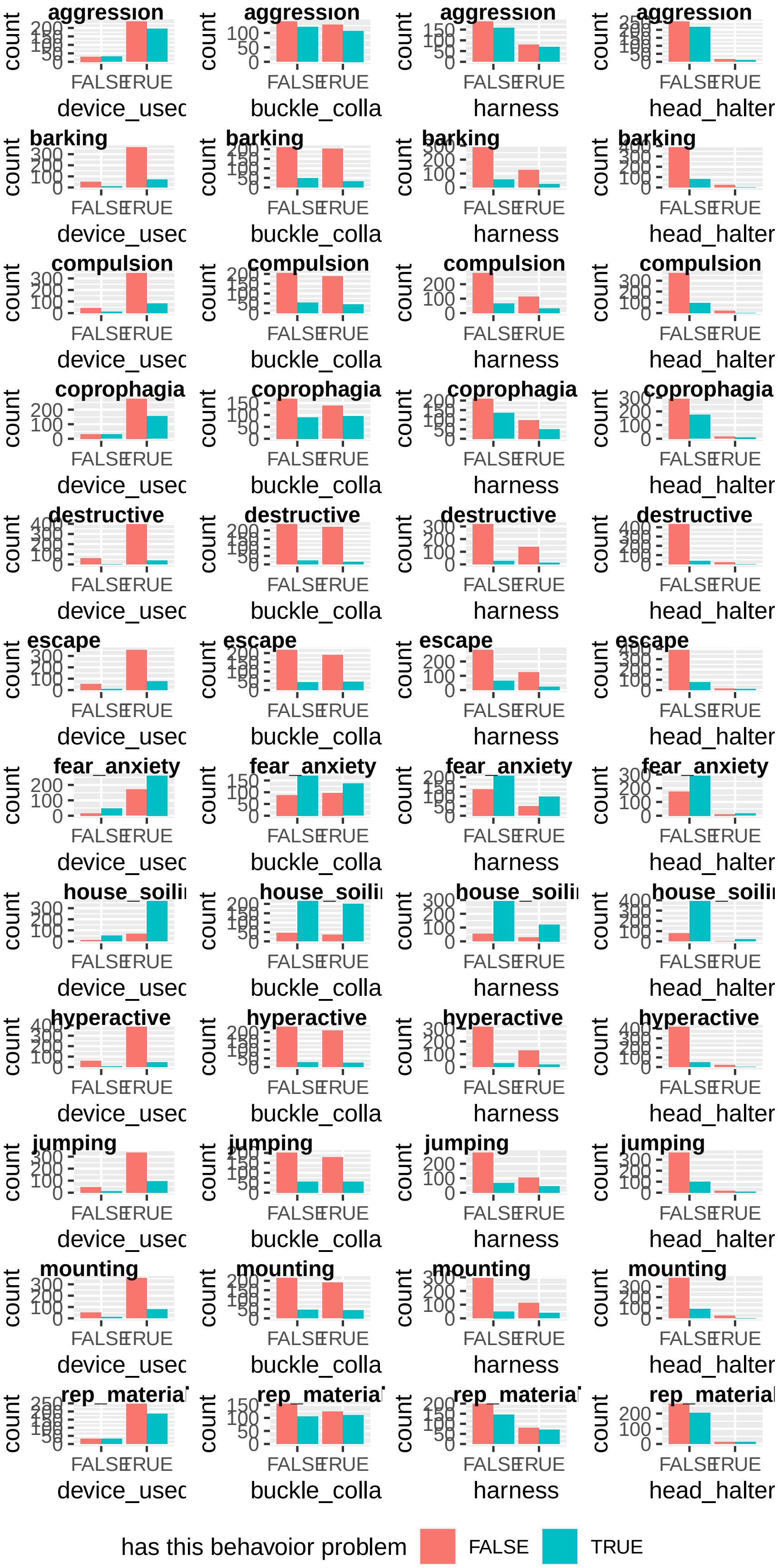
```r
attribs <- c(
  'train_1_3_mo',
  'train_4_mo',
  'train_5_6_mo',
  'train_start_age',
  'adj_train_technique'
)
attribs <- sort(attribs)

plot_list <- list()
cnt <- 1
labels <- NULL
for (i in 1:length(outcomes)) {
  outcome <- outcomes[i]
  for (j in 1:length(attribs)) {
    attrib <- attribs[j]
    p <- df %>%
      drop_na(attrib) %>%
      select(attrib, outcome) %>%
      ggplot(aes_string(x=attrib, fill=outcome)) +
      geom_bar(position = position_dodge(0.9)) +
      labs(fill='has this behavoior problem') +
      theme(legend.position='none')
    plot_list[[cnt]] <- p
    cnt <- cnt + 1
    labels <- c(labels, outcome)
  }
}
ggarrange(plotlist=plot_list, ncol=5, nrow=12, common.legend=TRUE,
          font.label=list(size=10), vjust=0.75, legend='bottom', labels=labels)
```



Next, we try to visual the impact

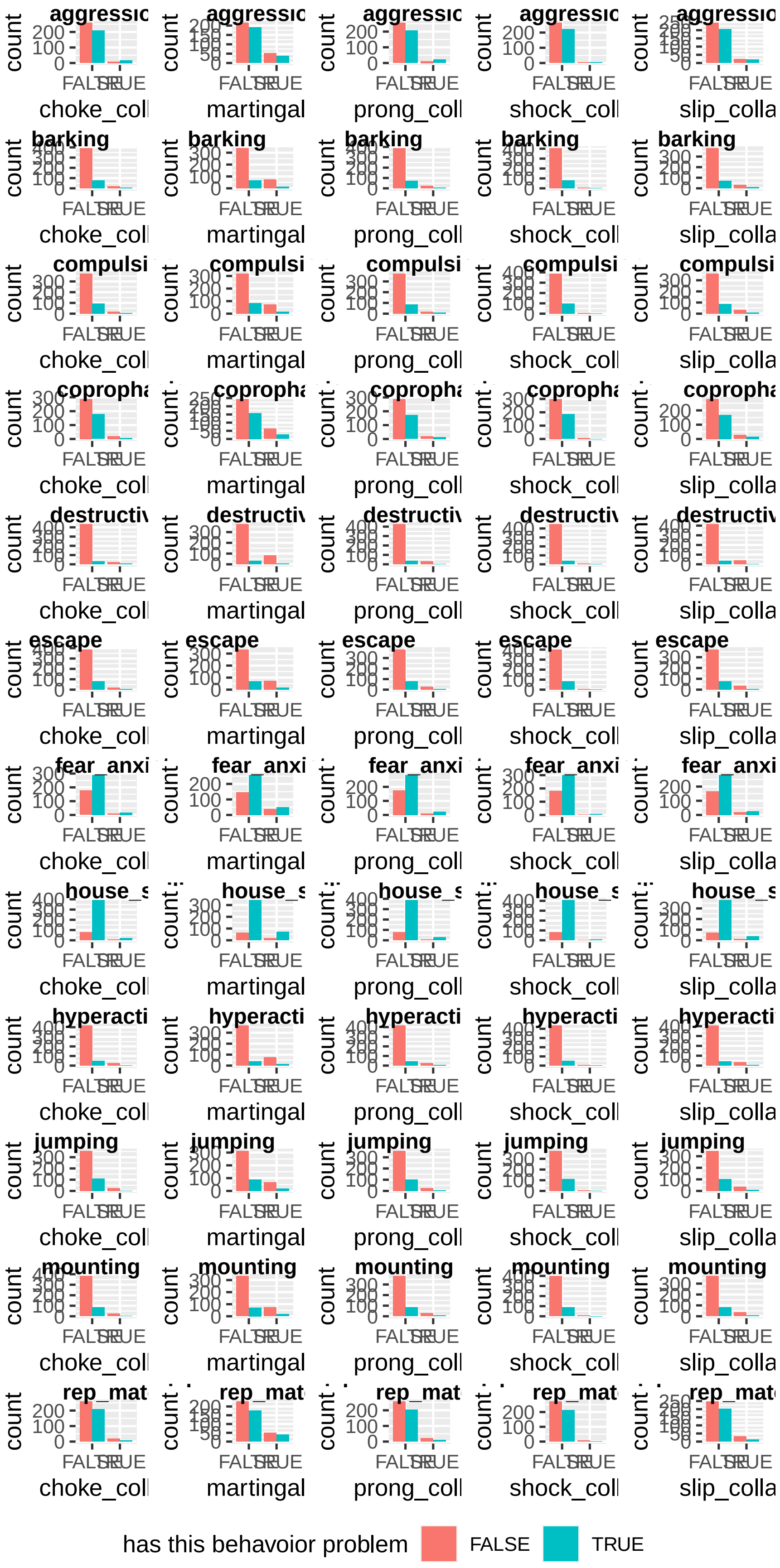
(if any) of a non-punishing restraining device choice. We also include the overall `device_used` column to see if there is a trend observed for restraining devices as a whole.

```
attribs <- c(
  'device_used',
  'buckle_collar',
  'harness',
  'head_halter'
)

plot_list <- list()
cnt <- 1
labels <- NULL
for (i in 1:length(outcomes)) {
  outcome <- outcomes[i]
  for (j in 1:length(attribs)) {
    attrib <- attribs[j]
    p <- df %>%
      drop_na(attrib) %>%
      select(attrib, outcome) %>%
      ggplot(aes_string(x=attrib, fill=outcome)) +
      geom_bar(position = position_dodge(0.9)) +
      labs(fill='has this behavoior problem') +
      theme(legend.position='none')
    plot_list[[cnt]] <- p
    cnt <- cnt + 1
    labels <- c(labels, outcome)
  }
}
ggarrange(plotlist=plot_list, ncol=4, nrow=12, common.legend=TRUE,
          font.label=list(size=10), vjust=0.75, legend='bottom', labels=labels)
```


aggression
barking
compulsion
coprophagia
destructive
escape
fear_anxiety
house_soili
hyperactive
jumping
mounting
rep_materia
count
FALSE
TRUE
device_usec
buckle_colla
harness
head_halter
has this behavoior problem
FALSE
TRUE

Last, we look at the impact of punishing restraining devices.

```r
attribs <- c(
  'martingale',
  'slip_collar',
  'shock_collar',
  'choke_collar',
  'prong_collar'
)
attribs <- sort(attribs)

plot_list <- list()
cnt <- 1
labels <- NULL
for (i in 1:length(outcomes)) {
  outcome <- outcomes[i]
  for (j in 1:length(attribs)) {
    attrib <- attribs[j]
    p <- df %>%
      drop_na(attrib) %>%
      select(attrib, outcome) %>%
      ggplot(aes_string(x=attrib, fill=outcome)) +
      geom_bar(position = position_dodge(0.9)) +
      labs(fill='has this behavoior problem') +
      theme(legend.position='none')
    plot_list[[cnt]] <- p
    cnt <- cnt + 1
    labels <- c(labels, outcome)
  }
}
ggarrange(plotlist=plot_list, ncol=5, nrow=12, common.legend=TRUE,
          font.label=list(size=10), vjust=0.75, legend='bottom', labels=labels)
```


aggressio
barking
compulsi
coprophä
destructiv
escape
fear_anxi
house_s
hyperacti
jumping
mounting
rep_mat
count
FALSE TRUE
choke_coll
martingal
prong_coll
shock_coll
slip_colla
has this behavoir problem
FALSE
TRUE