The DART Matrix

Daniel Spakowicz

4/24/2017

This is the start of Dan's effort on the final project for CBB750. The intention of this document is to perform some exploratory data analysis and to create summary slides for the presentation.

```
# Read in data
x <- read.csv("TheDartMatrix-deidentified.csv", as.is = TRUE)

# Read in variable key that defines classes
key <- read.csv("variable_key.csv")
levels(key$class)
```

```
## [1] "character" "factor"    "integer"   "numeric"
```

```
# Set classes
factors <- grep("factor", key$class)
char <- grep("character", key$class)
int <- grep("integer", key$class)
num <- grep("numeric", key$class)

x[,factors] <- lapply(x[factors], factor)
x[,char] <- lapply(x[char], as.character)
x[,int] <- lapply(x[int], as.integer)
x[,num] <- lapply(x[num], as.numeric)
```

Let's look at a quick table of when `fits_matrix` is called and when the DART consult is called.

```
table(x$fits_matrix)
```

0 1 322 353

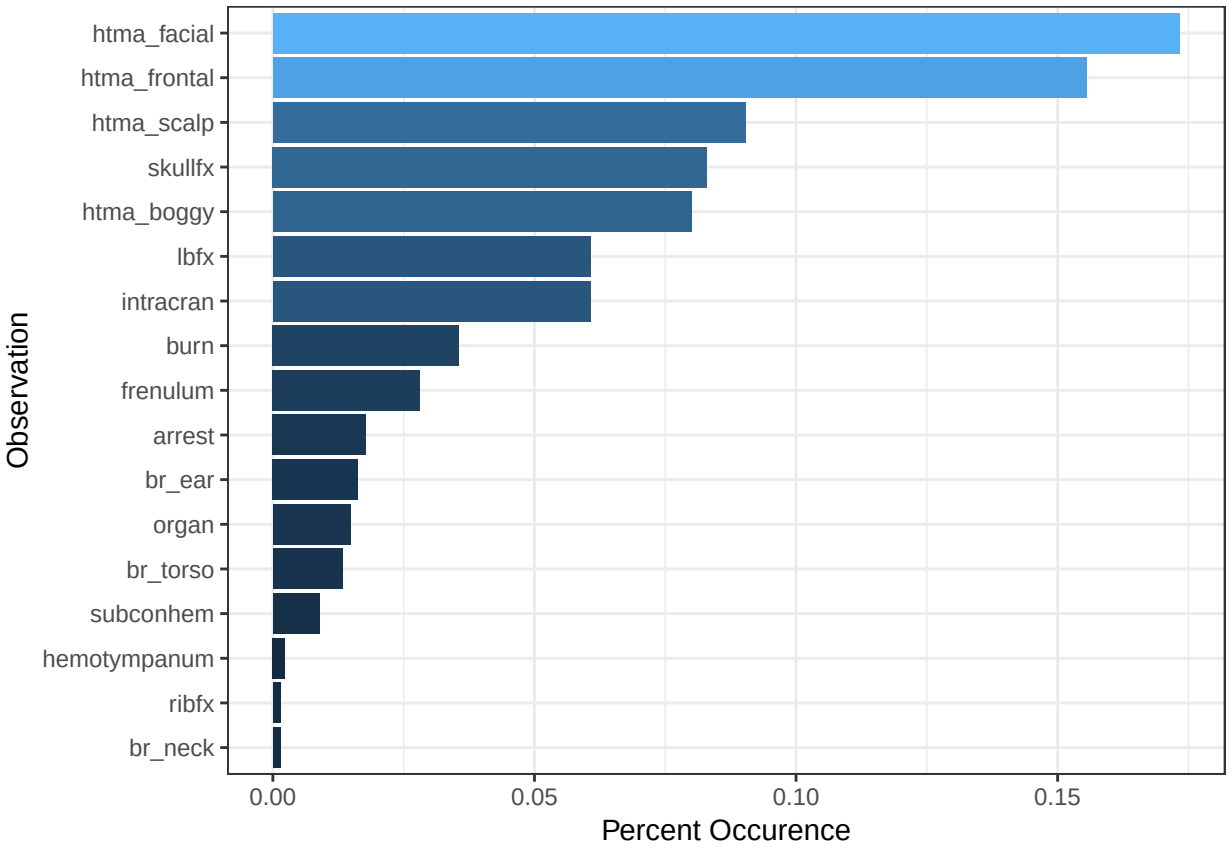
```
print(xtable(table(fits_matrix = x$fits_matrix, consult_dart = x$consult_dart)),
      comment = FALSE)
```

	0	1	2	3
0	312	9	1	0
1	221	122	0	10

I don't understand why `consult_dart` is 0-3.

Now I'll look at the frequency of the predictor variables.

```
mat <- data.frame(lapply(x[,5:21], as.numeric))
means <- data.frame(lapply(mat, function(x) mean(x-1, na.rm = TRUE)))
tmeans <- data.frame(var = names(means), perc_occurrence = t(means))
ggplot(tmeans, aes(x = reorder(var, perc_occurrence), y = perc_occurrence)) +
  geom_bar(stat = "identity", aes(fill = perc_occurrence)) +
  coord_flip() +
  labs(y = "Percent Occurence",
       x = "Observation") +
  theme_bw() +
  theme(legend.position = "none")
```


htma_facial
htma_frontal
htma_scalp
skullfx
htma_boggy
lbfx
intracran
burn
frenulum
arrest
br_ear
organ
br_torso
subconhem
hemotympanum
ribfx
br_neck
Observation
0.00
0.05
0.10
0.15
Percent Occurence