# Textfiles.com Analysis

## Preparation

```
# install stuff that we need later
if (!require("DT")) install.packages('DT')
if (!require("ggplot2")) install.packages('ggplot2')
if (!require("tidyverse")) install.packages('tidyverse')
if (!require("hrbrthemes")) install.packages('hrbrthemes')
if (!require("dplyr")) install.packages('dplyr')

# Load stuff we need later
library(readr)
library(DT)
library(ggplot2)
library(tidyverse)
library(hrbrthemes)
library(dplyr)
library(scales)

# and set the working directory
setwd("~/projects/bbs-for-independence/03_workspace")
```

## Import Data

```
# Read dataset summary from csv
dataset <- read_csv("./models/dataset.csv", show_col_types = FALSE)
dataset$charratioDelta = dataset$charratioB - dataset$charratioA
```

## Prepare Data

```
# Check the average of length, length_raw, avgcolumnsize, charratioA and charratioB
df = aggregate(x = dataset[, c(4,5,6,7,8,13)],
               by = list(dataset$category),
               FUN = function(x) list(
                 mean = round(mean(suppressWarnings(as.numeric(as.character(x))), na.rm=TRUE), digits = 2),
                 n = length(x)))
df <- do.call(data.frame, df) # bind columns which contain matrices back into the data frame
df <- as.data.frame(lapply(df, unlist)) # convert lists back to vectors

f_selection <- dataset %>% filter(!category %in% c("fidonet-on-the-internet", "tap", "floppies",
                                                   "exhibits", "artifacts", "piracy", "art",
                                                   "magazines", "digest"))
f_magazins <- dataset %>% filter(category == "magazines")
f_digest <- dataset %>% filter(category == "digest")

fun_charratioB_selection <- Vectorize( function(x) { nrow(f_selection %>% filter(charratioB > x)) } )
fun_charratioB_magazins <- Vectorize( function(x) { nrow(f_magazins %>% filter(charratioB > x)) } )
fun_charratioB_digest <- Vectorize( function(x) { nrow(f_digest %>% filter(charratioB > x)) } )

data_fun <- data.frame(x = seq(0,1,0.01),
                       n = c(fun_charratioB_selection(seq(0,1,0.01)),
                                 fun_charratioB_magazins(seq(0,1,0.01)),
                                 fun_charratioB_digest(seq(0,1,0.01))),
                       categories = rep(c("selection", "magazins", "digest"), each = 101))
```

# Summarize Data

```
cat("Anzahl Dateien: ", nrow(dataset))
```

```
## Anzahl Dateien:  105470
```

```
cat("Anzahl Kategorien: ", nrow(df))
```

```
## Anzahl Kategorien:  48
```

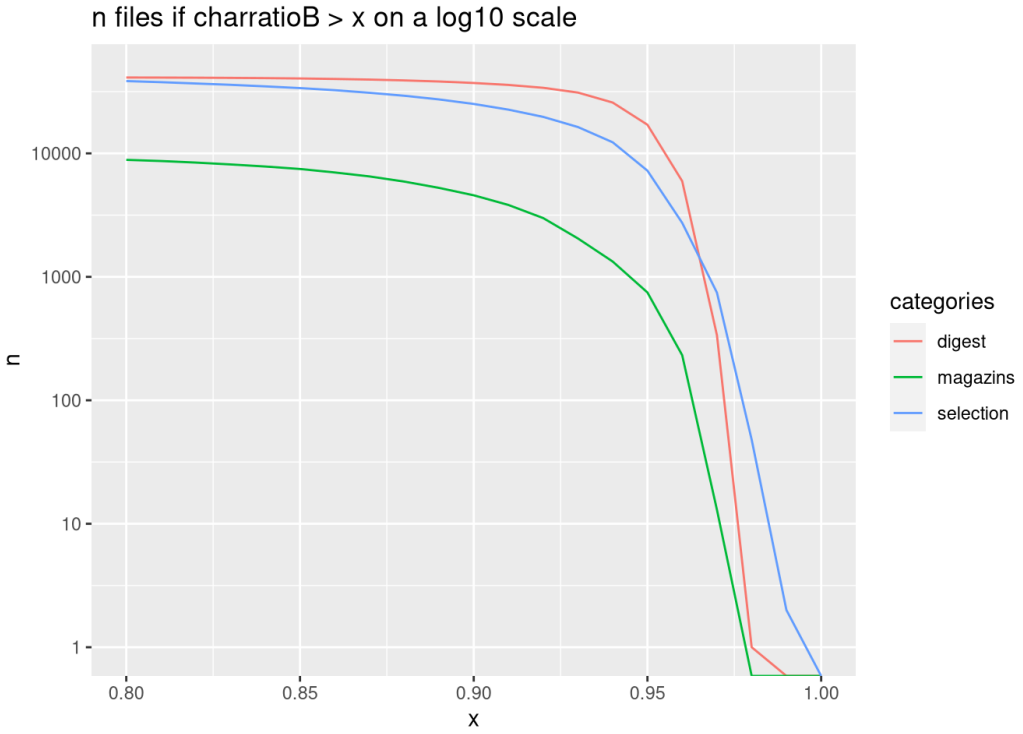
```
cat("Anzahl Dateien bei Filterung von tap, art, floppies, piracy, exhibits, magazines, digest:", nrow(f_selecti
on))
```

```
## Anzahl Dateien bei Filterung von tap, art, floppies, piracy, exhibits, magazines, digest: 44833
```

```
cat("Anzahl Dateien bei zusätzlicher Filterung von charratioB >0.95:", fun_charratioB_selection(0.95))
```

```
## Anzahl Dateien bei zusätzlicher Filterung von charratioB >0.95: 7242
```

```
# draw curve for fun_charratioB
ggplot(data_fun, aes(x, n, col = categories)) +
  geom_line() +
  xlim(0.8, 1) +
  scale_y_continuous(trans = log10_trans()) +
  ggtitle("n files if charratioB > x on a log10 scale")
```




```
datatable(df %>%
  arrange(desc(charratioB.mean)) %>%
  select(Group.1, length.n, length.mean, length_raw.mean, avgcolumnsize.mean,
         charratioA.mean, charratioB.mean, charratioDelta.mean),
  options = list(
    pageLength = 50,
    initComplete = JS("function(settings, json) {
                      $(this.api().table().header()).css({'font-size' : '12px'});
                      $('table.dataTable thead th').css({'padding' : '10px 18px 10px 0px'});
                      }")
    )
  )
```

Show 50 entries

Search:

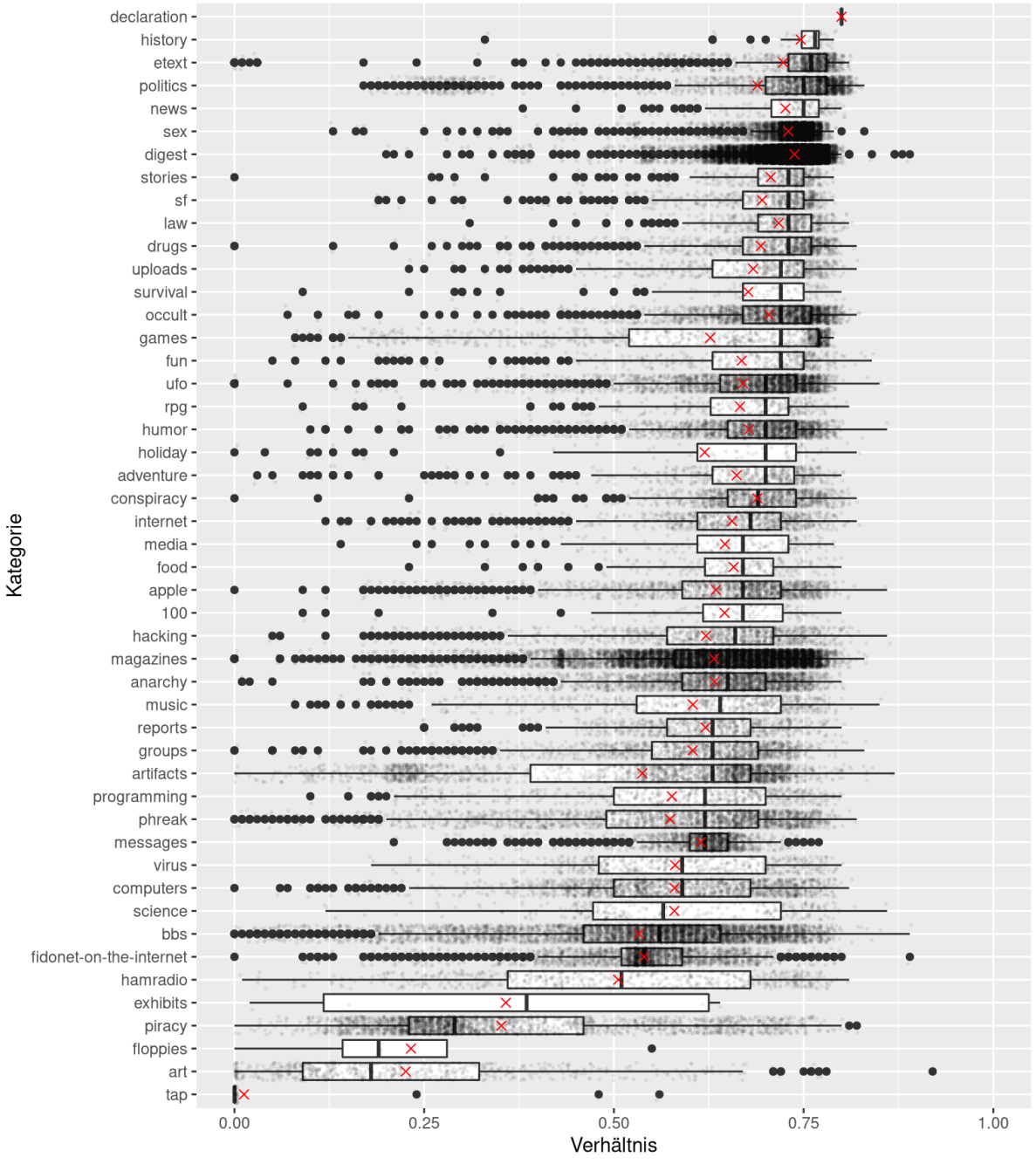
| | Group.1 | length.n | length.mean | length_raw.mean | avgcolumnsize.mean | charratioA.mean | charratioB.mean | charratioDelta.mean |
|---|---|---|---|---|---|---|---|---|
| 1 | declaration | 1 | 5089 | 5089 | 252.7 | 0.8 | 0.99 | 0.19 |
| 2 | history | 52 | 18230.58 | 18278.6 | 67.65 | 0.75 | 0.96 | 0.21 |
| 3 | sex | 5226 | 25250.56 | 25696.85 | 66.28 | 0.73 | 0.95 | 0.22 |
| 4 | digest | 41958 | 34516.91 | 34540.33 | 58 | 0.74 | 0.94 | 0.2 |
| 5 | etext | 1167 | 305073.34 | 307896.73 | 54.88 | 0.72 | 0.94 | 0.22 |
| 6 | law | 533 | 29662.52 | 30135.91 | 61.38 | 0.72 | 0.94 | 0.22 |
| 7 | politics | 2194 | 27928.15 | 28268.88 | 61.01 | 0.69 | 0.94 | 0.25 |
| 8 | stories | 474 | 27103.32 | 27541.75 | 63.82 | 0.71 | 0.94 | 0.23 |
| 9 | news | 184 | 12877.33 | 13061.82 | 78.99 | 0.73 | 0.93 | 0.2 |
| 10 | occult | 2400 | 28027.67 | 28461.17 | 61.07 | 0.7 | 0.92 | 0.22 |
| 11 | sf | 633 | 38383.36 | 39023.21 | 55.81 | 0.7 | 0.92 | 0.22 |
| 12 | survival | 105 | 16288.03 | 16550.34 | 62.77 | 0.68 | 0.92 | 0.25 |
| 13 | drugs | 1047 | 17824.46 | 17985.24 | 58.56 | 0.69 | 0.91 | 0.22 |
| 14 | fun | 430 | 25410.16 | 25883.55 | 58.23 | 0.67 | 0.91 | 0.24 |
| 15 | uploads | 560 | 8668.19 | 8697.51 | 95.72 | 0.68 | 0.91 | 0.23 |
| 16 | adventure | 550 | 11923.53 | 12104.18 | 66.53 | 0.66 | 0.9 | 0.24 |
| 17 | apple | 1553 | 17911 | 18129.07 | 71.52 | 0.63 | 0.9 | 0.26 |
| 18 | conspiracy | 1111 | 21273.07 | 21478.9 | 58.46 | 0.69 | 0.9 | 0.21 |
| 19 | food | 213 | 9774.92 | 9995.34 | 48.26 | 0.66 | 0.9 | 0.24 |
| 20 | humor | 2061 | 13307.51 | 13585.09 | 52.6 | 0.68 | 0.9 | 0.22 |
| 21 | rpg | 300 | 41099.01 | 41704.22 | 56.14 | 0.67 | 0.9 | 0.24 |
| 22 | anarchy | 2509 | 14516.98 | 14765.6 | 62.18 | 0.63 | 0.89 | 0.25 |
| 23 | media | 164 | 39436.89 | 40091.92 | 51.39 | 0.65 | 0.89 | 0.24 |
| 24 | ufo | 2928 | 12439.76 | 12679.46 | 60.37 | 0.67 | 0.89 | 0.22 |
| 25 | 100 | 100 | 28158.48 | 28532.05 | 56.12 | 0.65 | 0.88 | 0.24 |
| 26 | games | 980 | 18863.05 | 19138.54 | 60.01 | 0.63 | 0.88 | 0.25 |
| 27 | internet | 849 | 44187.19 | 44850.73 | 54.01 | 0.66 | 0.88 | 0.22 |
| 28 | groups | 1269 | 13385.23 | 13555.14 | 125.96 | 0.6 | 0.87 | 0.27 |
| 29 | hacking | 1107 | 28180.3 | 28623.02 | 62.17 | 0.62 | 0.87 | 0.25 |
| 30 | magazines | 10630 | 27312.95 | 27550.01 | 121.91 | 0.63 | 0.87 | 0.23 |
| 31 | music | 608 | 25020.44 | 25376.26 | 51.51 | 0.6 | 0.86 | 0.26 |
| 32 | reports | 713 | 11362.56 | 11559.32 | 67.8 | 0.62 | 0.86 | 0.24 |
| 33 | virus | 545 | 14449.13 | 14753.63 | 57.81 | 0.58 | 0.86 | 0.28 |
| 34 | programming | 601 | 37755.52 | 38504 | 55.11 | 0.58 | 0.85 | 0.28 |

| | Group.1 | length.n | length.mean | length_raw.mean | avgcolumnsize.mean | charratioA.mean | charratioB.mean | charratioDelta.mean |
|---|---|---|---|---|---|---|---|---|
| 35 | computers | 1699 | 22515.02 | 22964.09 | 55.69 | 0.58 | 0.84 | 0.26 |
| 36 | holiday | 73 | 5081.51 | 5191.74 | 140.96 | 0.62 | 0.84 | 0.23 |
| 37 | phreak | 2195 | 15278.46 | 15573.29 | 59.61 | 0.57 | 0.84 | 0.26 |
| 38 | messages | 1543 | 41495.49 | 42262.34 | 50.03 | 0.61 | 0.83 | 0.21 |
| 39 | bbs | 5242 | 23949.07 | 24374.74 | 62.4 | 0.53 | 0.81 | 0.28 |
| 40 | hamradio | 636 | 14112.25 | 14384.66 | 55 | 0.51 | 0.81 | 0.31 |
| 41 | science | 278 | 18247 | 18479.9 | 57.43 | 0.58 | 0.81 | 0.23 |
| 42 | fidonet-on-the-internet | 2498 | 180826.98 | 182547.84 | 49.15 | 0.54 | 0.78 | 0.24 |
| 43 | art | 656 | 24714.97 | 24756.06 | 86.08 | 0.23 | 0.74 | 0.51 |
| 44 | piracy | 2539 | 9177.07 | 9326.82 | 85.12 | 0.35 | 0.73 | 0.38 |
| 45 | artifacts | 2246 | 37027.04 | 37305.74 | 69.53 | 0.54 | 0.7 | 0.16 |
| 46 | exhibits | 4 | 97722.25 | 97724.75 | 119.16 | 0.36 | 0.54 | 0.18 |
| 47 | floppies | 4 | 8654 | 8654 | 29.81 | 0.23 | 0.31 | 0.08 |
| 48 | tap | 102 | 853.44 | 877 | 1.04 | 0.01 | 0.02 | 0.01 |

# Plots

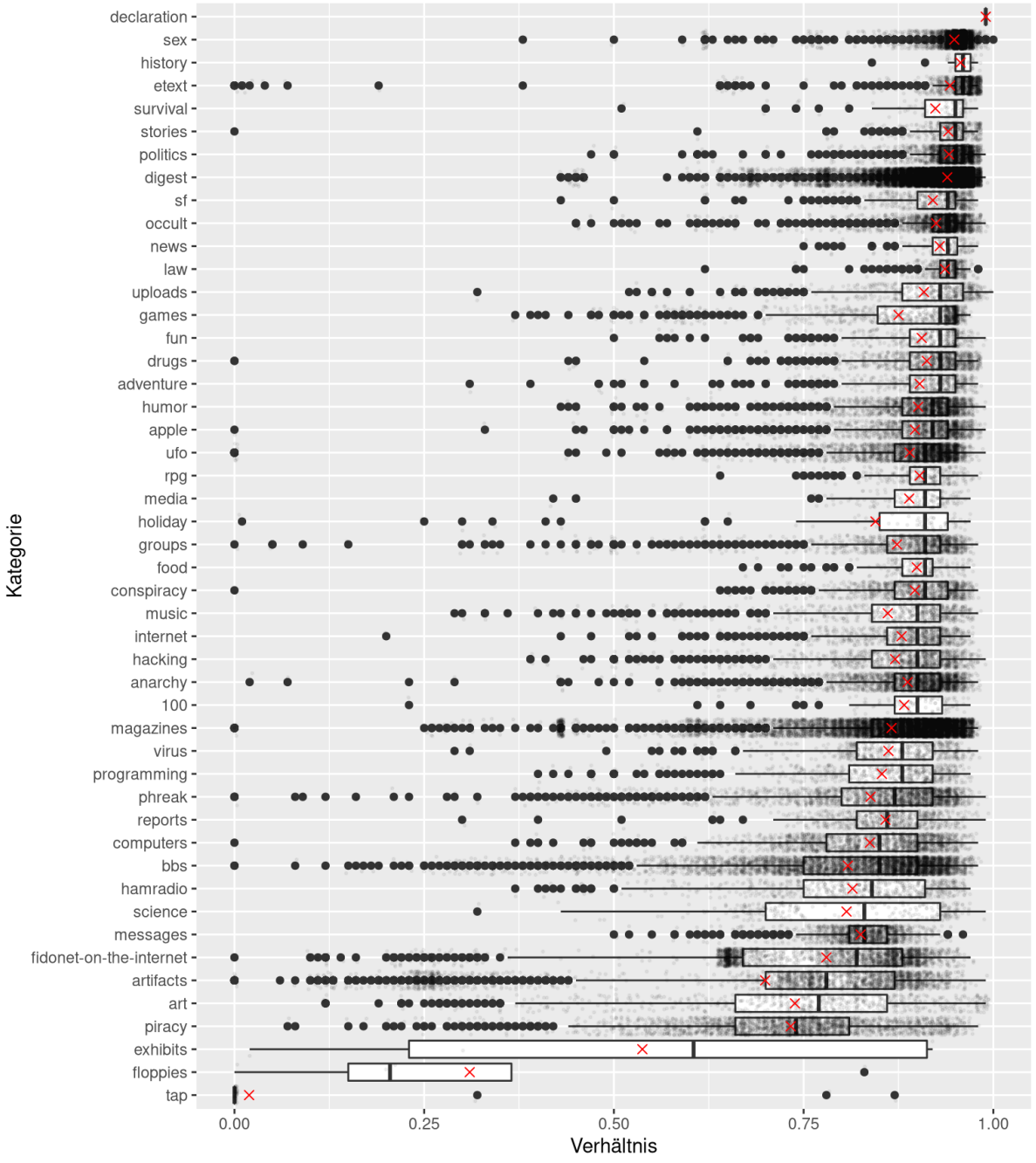
## Verhältnis Text (exkl. Satz- und Leerzeichen) zu Dateilänge

```
dataset %>%
  ggplot( aes(x=reorder(category, charratioA, FUN = median),
              y=charratioA, group=category)) +
    geom_boxplot() +
    theme(
      legend.position="none",
      plot.title = element_text(size=11)
    ) +
    geom_jitter(color="black", size=0.4, alpha=0.05) +
    stat_summary(fun.y=mean, geom="point", shape=4, size=2, color="red", fill="red") +
    coord_flip() +
    ylim(0, 1) +
    xlab("Kategorie") +
    ylab("Verhältnis")
```



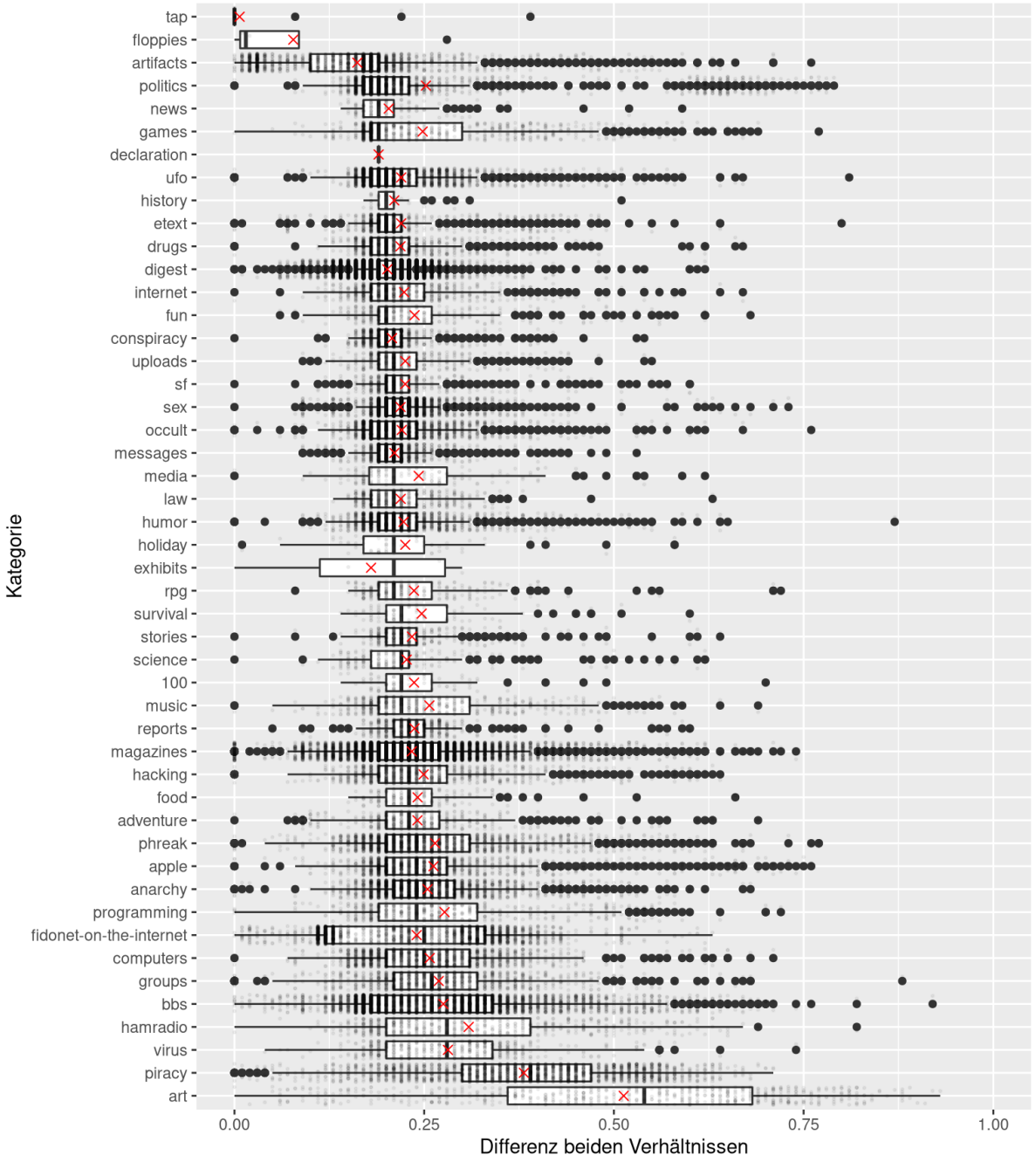
# Verhältnis Text (inkl. Satz- und Leerzeichen) zu Dateilänge

```
# create plot: charratioB
dataset %>%
  ggplot( aes(x=reorder(category, charratioB, FUN = median),
              y=charratioB, group=category)) +
    geom_boxplot() +
    theme(
      legend.position="none",
      plot.title = element_text(size=11)
    ) +
    geom_jitter(color="black", size=0.4, alpha=0.05) +
    stat_summary(fun.y=mean, geom="point", shape=4, size=2, color="red", fill="red") +
    coord_flip() +
    ylim(0, 1) +
    xlab("Kategorie") +
    ylab("Verhältnis")
```



# Differenz beiden Verhältnissen (inkl. minus exkl. Satz- und Leerzeichen zu Dateilänge)

```
dataset %>%
  ggplot( aes(x=reorder(category, charratioA-charratioB, FUN = median),
              y=charratioB-charratioA, group=category)) +
    geom_boxplot() +
    theme(
      legend.position="none",
      plot.title = element_text(size=11)
    ) +
    geom_jitter(color="black", size=0.4, alpha=0.05) +
    stat_summary(fun.y=mean, geom="point", shape=4, size=2, color="red", fill="red") +
    coord_flip() +
    ylim(0, 1) +
    xlab("Kategorie") +
    ylab("Differenz beiden Verhältnissen")
```



# Apply filtering and extend filtering

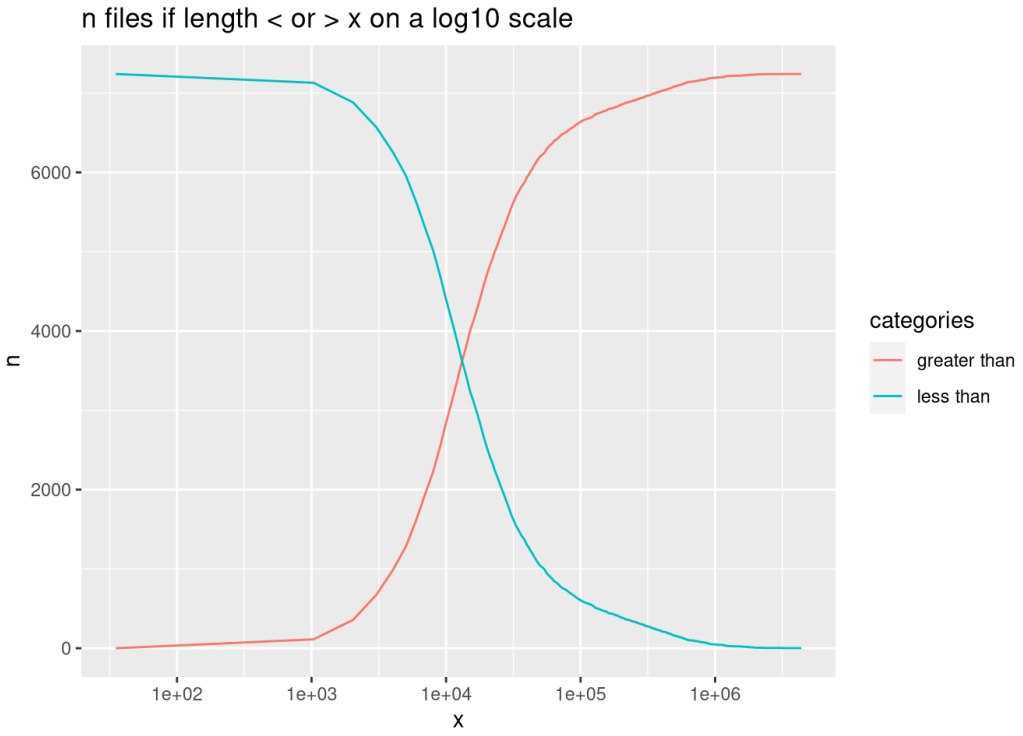
```
data_names_exclude <- c("fidonet-on-the-internet","tap","floppies","exhibits",
                        "artifacts","piracy", "art", "magazines", "digest")

dataset_filtered = dataset %>%
  filter(!category %in% data_names_exclude) %>%
  filter(charratioB > 0.95)

fun_length_selection_lt <- Vectorize( function(x) { nrow(dataset_filtered %>% filter(length < x)) } )
fun_length_selection_gt <- Vectorize( function(x) { nrow(dataset_filtered %>% filter(length > x)) } )

length_fun_seq = seq(min(dataset_filtered$length),max(dataset_filtered$length),1000)
length_fun <- data.frame(x = length_fun_seq,
                         n = c(fun_length_selection_lt(length_fun_seq),
                               fun_length_selection_gt(length_fun_seq)),
                         categories = rep(c("greater than", "less than"), each = length(length_fun_seq)))

ggplot(length_fun, aes(x, n, col = categories)) +
  geom_line() +
  scale_x_continuous(trans = log10_trans()) +
  ggtitle("n files if length < or > x on a log10 scale")
```




```
dataset_filtered_2 = dataset %>%
  filter(!category %in% data_names_exclude) %>%
  filter(charratioB > 0.95) %>%
  filter(length > 300) %>%
  filter(length < 30000)

cat("Anzahl Dateien mit gefilterter Länge: ", nrow(dataset_filtered_2))
```

```
## Anzahl Dateien mit gefilterter Länge:  5510
```

```
cat("Gesammt Länge der Dateien mit gefilterter Länge: ", sum(dataset_filtered_2$length),
    "≙",round(sum(dataset_filtered_2$length)/1000000),"MB ")
```

```
## Gesammt Länge der Dateien mit gefilterter Länge:  62128188 ≙ 62 MB
```

```r
## Determine highlighted regions
v <- rep(0, length(length_fun_seq))
v[c(round(300/1000):round(30000/1000))] <- 1

## Get the start and end points for highlighted regions
inds <- diff(c(0, v))
start <- length_fun$x[inds == 1]
end <- length_fun$x[inds == -1]
if (length(start) > length(end)) end <- c(end, tail(length_fun$x, 1))

## highlight region data
rects <- data.frame(start=start, end=end, group=seq_along(start))

ggplot(length_fun, aes(x, n, col = categories)) +
  geom_line() +
  scale_x_continuous(trans = log10_trans()) +
  geom_rect(data=rects, inherit.aes=FALSE, aes(xmin=300, xmax=30000, ymin=min(length_fun$n),
                ymax=max(length_fun$n), group=group), color="black", fill="transparent", alpha=0) +
  ggtitle("n files if length < or > x on a log10 scale with range")
```