EFA

```r
set.seed(42)

library(igraph)
```

```
##
## Attaching package: 'igraph'

## The following objects are masked from 'package:stats':
##
##     decompose, spectrum

## The following object is masked from 'package:base':
##
##     union
```

```r
library(moments) # for skewness()
library(robustbase)
library(QuantPsyc) # for the multivariate normality test
```

```
## Loading required package: boot

##
## Attaching package: 'boot'

## The following object is masked from 'package:robustbase':
##
##     salinity

## Loading required package: dplyr

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:igraph':
##
##     as_data_frame, groups, union

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## Loading required package: purrr

##
## Attaching package: 'purrr'

## The following objects are masked from 'package:igraph':
##
##     compose, simplify
## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

##
## Attaching package: 'QuantPsyc'

## The following object is masked from 'package:base':
##
##     norm
```

```r
library(dunn.test)
library(nFactors) # for the scree plot
```

```
## Loading required package: lattice

##
## Attaching package: 'lattice'

## The following object is masked from 'package:boot':
##
##     melanoma

##
## Attaching package: 'nFactors'

## The following object is masked from 'package:lattice':
##
##     parallel
```

```r
library(psych) # for PA FA
```

```
##
## Attaching package: 'psych'

## The following object is masked from 'package:boot':
##
##     logit
```

```r
library(caret) # highly correlated features removal
```

```
## Loading required package: ggplot2

##
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':
##
##     %+%, alpha

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v readr     2.1.5     v tidyr     1.3.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x lubridate::%--%()       masks igraph::%--%()
## x ggplot2::%+%()          masks psych::%+%()
## x ggplot2::alpha()        masks psych::alpha()
## x tibble::as_data_frame() masks dplyr::as_data_frame(), igraph::as_data_frame()
## x purrr::compose()        masks igraph::compose()
## x tidyr::crossing()       masks igraph::crossing()
## x dplyr::filter()         masks stats::filter()
## x dplyr::lag()            masks stats::lag()
## x caret::lift()           masks purrr::lift()
## x MASS::select()          masks dplyr::select()
## x purrr::simplify()       masks igraph::simplify()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(paletteer) # color palettes

library(conflicted) # to resolve QuantPsyc x dplyr conflicts
conflict_prefer("select", "dplyr")
```

```
## [conflicted] Will prefer dplyr::select over any other package.
```

```r
conflict_prefer("filter", "dplyr")
```

```
## [conflicted] Will prefer dplyr::filter over any other package.
```

Load and tidy data

```r
pretty_names <- read_csv("../feat_name_mapping.csv")
```

```
## Rows: 85 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): name_orig, name_pretty
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
prettify_feat_name <- function(x) {
  name <- pull(pretty_names %>%
    filter(name_orig == x), name_pretty)
  if (length(name) == 1) {
    return(name)
  } else {
    return(x)
  }
}

prettify_feat_name_vector <- function(x) {
  map(
    x,
    prettify_feat_name
  ) %>% unlist()
}

data <- read_csv("../measurements/measurements.csv")
```

```
## Rows: 754 Columns: 108
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (20): fpath, KUK_ID, FileName, FileFormat, FolderPath, subcorpus, Source...
## dbl (85): RuleAbstractNouns, RuleAmbiguousRegards, RuleAnaphoricReferences, ...
## lgl  (3): ClarityPursuit, SyllogismBased, Bindingness
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
.firstnonmetacolumn <- 17

data_no_nas <- data %>%
  select(!c(
    fpath,
    # KUK_ID,
    # FileName,
    FolderPath,
    # subcorpus,
    DocumentTitle,
    ClarityPursuit,
    Readability,
    SyllogismBased,
    SourceDB
  )) %>%
  # replace -1s in variation coefficients with NAs
  mutate(across(c(
    `RuleDoubleAdpos.max_allowable_distance.v`,
    `RuleTooManyNegations.max_negation_frac.v`,
    `RuleTooManyNegations.max_allowable_negations.v`,
    `RuleTooManyNominalConstructions.max_noun_frac.v`,
    `RuleTooManyNominalConstructions.max_allowable_nouns.v`,
    `RuleCaseRepetition.max_repetition_count.v`,
    `RuleCaseRepetition.max_repetition_frac.v`,
    `RulePredSubjDistance.max_distance.v`,
    `RulePredObjDistance.max_distance.v`,
    `RuleInfVerbDistance.max_distance.v`,
    `RuleMultiPartVerbs.max_distance.v`,
    `RuleLongSentences.max_length.v`,
    `RulePredAtClauseBeginning.max_order.v`,
    `mattr.v`,
    `maentropy.v`
  ), ~ na_if(.x, -1))) %>%
  # replace NAs with 0s
  replace_na(list(
    RuleGPcoordovs = 0,
    RuleGPdeverbaddr = 0,
    RuleGPpatinstr = 0,
    RuleGPdeverbsubj = 0,
    RuleGPadjective = 0,
    RuleGPpatbenperson = 0,
    RuleGPwordorder = 0,
    RuleDoubleAdpos = 0,
    RuleDoubleAdpos.max_allowable_distance = 0,
    RuleDoubleAdpos.max_allowable_distance.v = 0,
    RuleAmbiguousRegards = 0,
    RuleReflexivePassWithAnimSubj = 0,
    RuleTooManyNegations = 0,
    RuleTooManyNegations.max_negation_frac = 0,
    RuleTooManyNegations.max_negation_frac.v = 0,
    RuleTooManyNegations.max_allowable_negations = 0,
    RuleTooManyNegations.max_allowable_negations.v = 0,
    RuleTooManyNominalConstructions.max_noun_frac.v = 0,
    RuleTooManyNominalConstructions.max_allowable_nouns.v = 0,
    RuleFunctionWordRepetition = 0,
    RuleCaseRepetition.max_repetition_count.v = 0,
    RuleCaseRepetition.max_repetition_frac.v = 0,
    RuleWeakMeaningWords = 0,
    RuleAbstractNouns = 0,
    RuleRelativisticExpressions = 0,
    RuleConfirmationExpressions = 0,
    RuleRedundantExpressions = 0,
    RuleTooLongExpressions = 0,
    RuleAnaphoricReferences = 0,
    RuleLiteraryStyle = 0,
    RulePassive = 0,
    RulePredSubjDistance = 0,
    RulePredSubjDistance.max_distance = 0,
    RulePredSubjDistance.max_distance.v = 0,
    RulePredObjDistance = 0,
    RulePredObjDistance.max_distance = 0,
    RulePredObjDistance.max_distance.v = 0,
    RuleInfVerbDistance = 0,
    RuleInfVerbDistance.max_distance = 0,
    RuleInfVerbDistance.max_distance.v = 0,
    RuleMultiPartVerbs = 0,
    RuleMultiPartVerbs.max_distance = 0,
    RuleMultiPartVerbs.max_distance.v = 0,
    RuleLongSentences.max_length.v = 0,
    RulePredAtClauseBeginning.max_order.v = 0,
    RuleVerbalNouns = 0,
    RuleDoubleComparison = 0,
    RuleWrongValencyCase = 0,
    RuleWrongVerbonominalCase = 0,
    RuleIncompleteConjunction = 0
  ))
```

Outliers

```
data_no_out <- data_no_nas %>% filter(KUK_ID != "CzCDC_SupC088540")
```

Normalization

```
data_clean <- data_no_out %>%
  # norm data expected to correlate with text length
  mutate(across(c(
    RuleGPcoordovs,
    RuleGPdeverbaddr,
    RuleGPpatinstr,
    RuleGPdeverbsubj,
    RuleGPadjective,
    RuleGPpatbenperson,
    RuleGPwordorder,
    RuleDoubleAdpos,
    RuleAmbiguousRegards,
    RuleFunctionWordRepetition,
    RuleWeakMeaningWords,
    RuleAbstractNouns,
    RuleRelativisticExpressions,
    RuleConfirmationExpressions,
    RuleRedundantExpressions,
    RuleTooLongExpressions,
    RuleAnaphoricReferences,
    RuleLiteraryStyle,
    RulePassive,
    RuleVerbalNouns,
    RuleDoubleComparison,
    RuleWrongValencyCase,
    RuleWrongVerbonominalCase,
    RuleIncompleteConjunction,
    num_hapax,
    RuleReflexivePassWithAnimSubj,
    RuleTooManyNominalConstructions,
    RulePredSubjDistance,
    RuleMultiPartVerbs,
    RulePredAtClauseBeginning
  ), ~ .x / word_count)) %>%
  mutate(across(c(
    RuleTooFewVerbs,
    RuleTooManyNegations,
    RuleCaseRepetition,
    RuleLongSentences,
    RulePredObjDistance,
    RuleInfVerbDistance
  ), ~ .x / sent_count)) %>%
  # remove variables identified as "u counts"
  select(!c(
    RuleTooFewVerbs,
    RuleTooManyNegations,
    RuleTooManyNominalConstructions,
```

```r
    RuleCaseRepetition,
    RuleLongSentences,
    RulePredAtClauseBeginning,
    syllab_count,
    char_count
  )) %>%
  # remove variables identified as unreliable
  select(!c(
    RuleAmbiguousRegards,
    RuleFunctionWordRepetition,
    RuleDoubleComparison,
    RuleWrongValencyCase,
    RuleWrongVerbonominalCase
  )) %>%
  # remove artificially limited variables
  select(!c(
    RuleCaseRepetition.max_repetition_frac,
    RuleCaseRepetition.max_repetition_frac.v
  )) %>%
  # remove further variables belonging to the 'acceptability' category
  select(!c(RuleIncompleteConjunction)) %>%
  mutate(across(c(
    class,
    FileFormat,
    subcorpus,
    DocumentVersion,
    LegalActType,
    Objectivity,
    AuthorType,
    RecipientType,
    RecipientIndividuation,
    Anonymized
  ), ~ as.factor(.x)))

# no NAs should be present now
data_clean[!complete.cases(data_clean), ]
```

```
## # A tibble: 753 x 85
##    KUK_ID                  FileName FileFormat subcorpus SourceID DocumentVersion
##    <chr>                   <chr>    <fct>      <fct>     <chr>    <fct>
##  1 673b7a37c6537d54ff062~ 002_Kom~ TXT        KUKY      <NA>     Original
##  2 673b7a37c6537d54ff062~ 006_Chc~ TXT        KUKY      <NA>     Redesign
##  3 673b7a37c6537d54ff062~ 004_Nev~ TXT        KUKY      <NA>     Original
##  4 673b7a37c6537d54ff062~ 008_Pol~ TXT        KUKY      <NA>     Original
##  5 673b7a37c6537d54ff062~ 005_Och~ TXT        KUKY      <NA>     Original
##  6 673b7a37c6537d54ff062~ 016_Obc~ TXT        KUKY      <NA>     Original
##  7 673b7a37c6537d54ff062~ 019_Dět~ TXT        KUKY      <NA>     Redesign
##  8 673b7a37c6537d54ff062~ 007_DŮC~ TXT        KUKY      <NA>     Redesign
##  9 673b7a37c6537d54ff062~ 024_Opa~ TXT        KUKY      <NA>     Original
## 10 673b7a37c6537d54ff062~ 047_Dav~ TXT        KUKY      <NA>     Original
## # i 743 more rows
## # i 79 more variables: ParentDocumentID <chr>, LegalActType <fct>,
## #   Objectivity <fct>, Bindingness <lgl>, AuthorType <fct>,
## #   RecipientType <fct>, RecipientIndividuation <fct>, Anonymized <fct>,
## #   `Recipient Type` <chr>, class <fct>, RuleAbstractNouns <dbl>,
## #   RuleAnaphoricReferences <dbl>,
## #   RuleCaseRepetition.max_repetition_count <dbl>, ...
```

```
data_clean_scaled <- data_clean %>%
  mutate(across(class, ~ .x == "good")) %>%
  mutate(across(.firstnonmetacolumn:length(names(data_clean)), ~ scale(.x)))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `across(.firstnonmetacolumn:length(names(data_clean)),
##   ~scale(.x))`.
## Caused by warning:
## ! Using an external vector in selections was deprecated in tidyselect 1.1.0.
## i Please use `all_of()` or `any_of()` instead.
##   # Was:
##   data %>% select(.firstnonmetacolumn)
##
##   # Now:
##   data %>% select(all_of(.firstnonmetacolumn))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
```

Important features identification

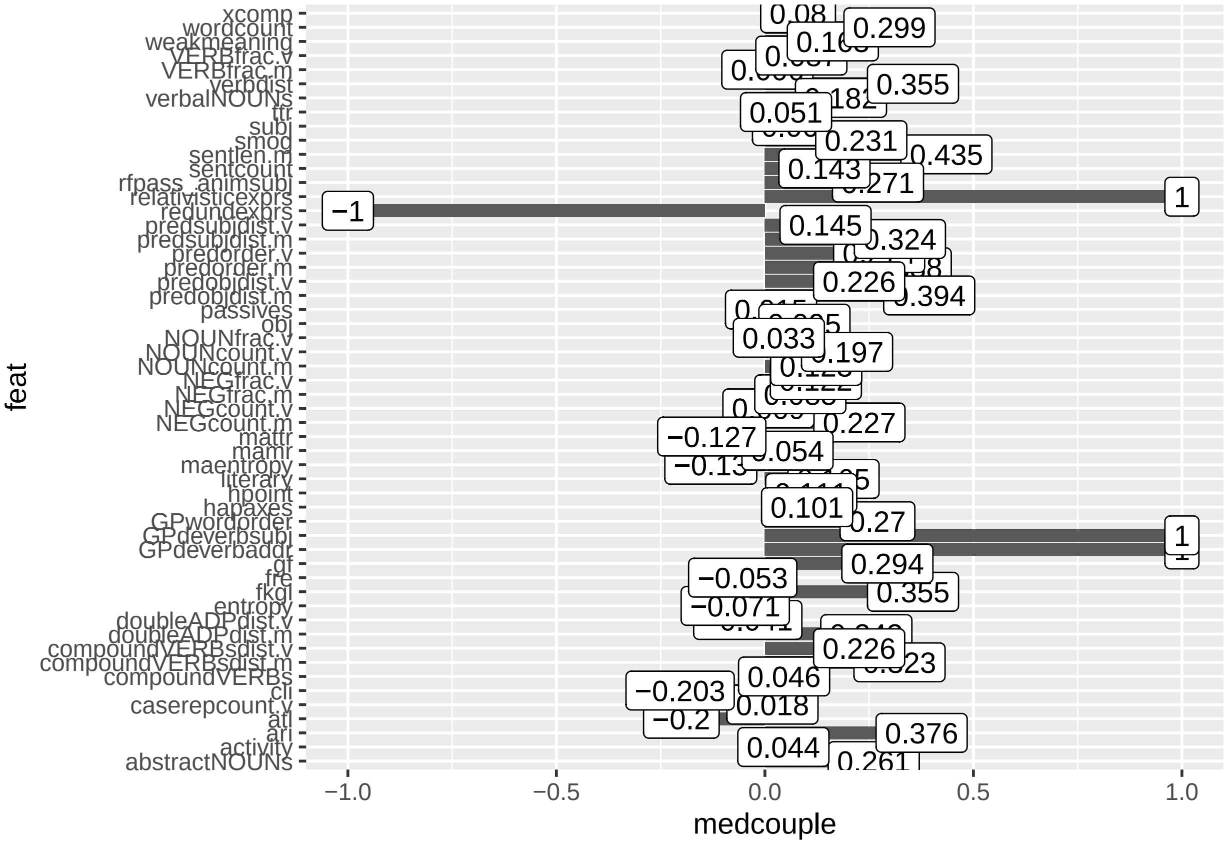
```
feature_importances <- tibble(
  feat_name = character(), p_value = numeric()
)

for (i in .firstnonmetacolumn:ncol(data_clean)) {
  fname <- names(data_clean)[i]

  formula_single <- reformulate(fname, "class")

  glm_model <- glm(formula_single, data_clean, family = "binomial")
  glm_coefficients <- summary(glm_model)$coefficients
  row_index <- which(rownames(glm_coefficients) == fname)
  p_value <- glm_coefficients[row_index, 4]

  feature_importances <- feature_importances %>%
    add_row(feat_name = fname, p_value = p_value)
}
feature_importances
```

```
## # A tibble: 69 x 2
##    feat_name                                  p_value
##    <chr>                                        <dbl>
##  1 RuleAbstractNouns                         0.00206
##  2 RuleAnaphoricReferences                   0.673
##  3 RuleCaseRepetition.max_repetition_count   0.0645
##  4 RuleCaseRepetition.max_repetition_count.v 0.00443
##  5 RuleConfirmationExpressions               0.0960
##  6 RuleDoubleAdpos                           0.316
##  7 RuleDoubleAdpos.max_allowable_distance    0.000161
##  8 RuleDoubleAdpos.max_allowable_distance.v  0.00000276
##  9 RuleGPadjective                            0.381
## 10 RuleGPcoordovs                             0.838
## # i 59 more rows
```

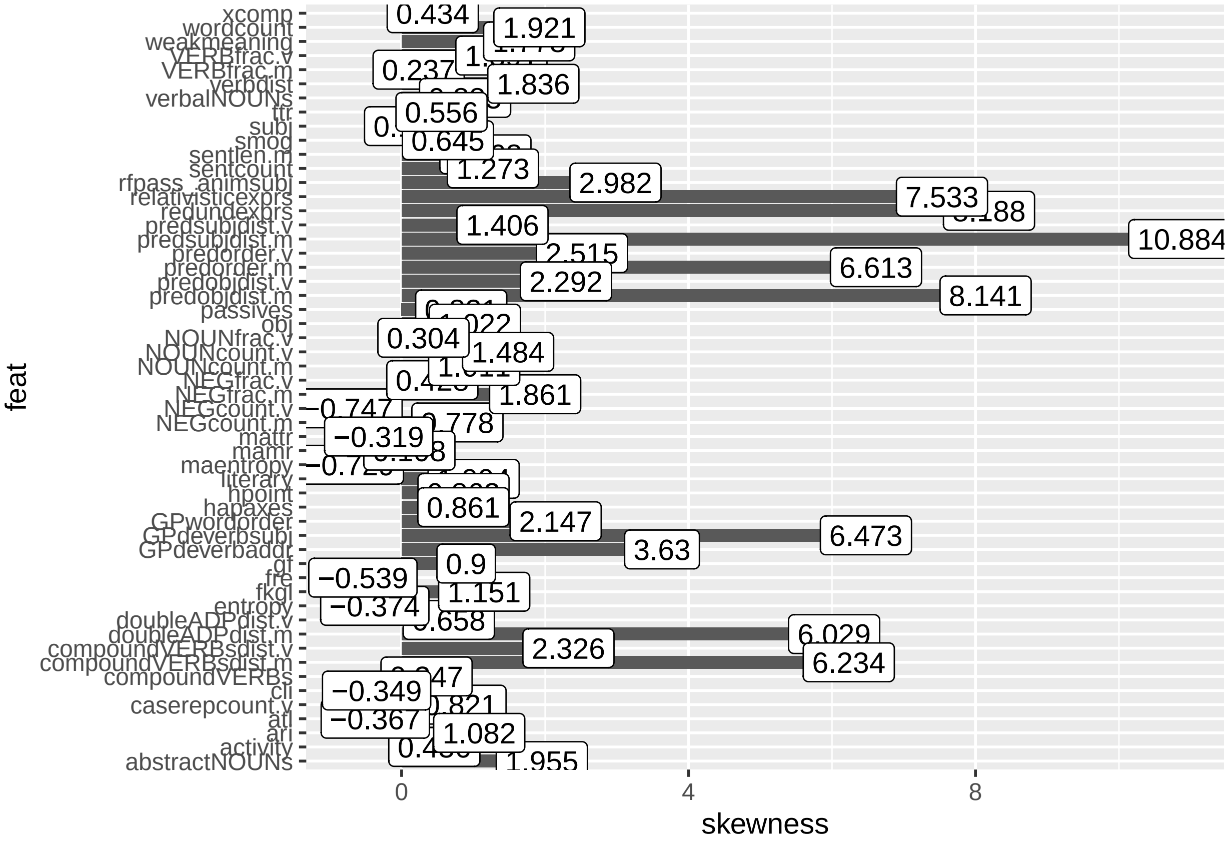
```r
selected_features <- feature_importances %>%
  mutate(selected = p_value <= 0.05)
selected_features %>% write_csv("selected_features.csv")
selected_features_names <- selected_features %>%
  filter(selected) %>%
  pull(feat_name)

data_purish <- data_clean %>% select(any_of(selected_features_names))
colnames(data_purish) <- prettify_feat_name_vector(colnames(data_purish))
```

Skewness

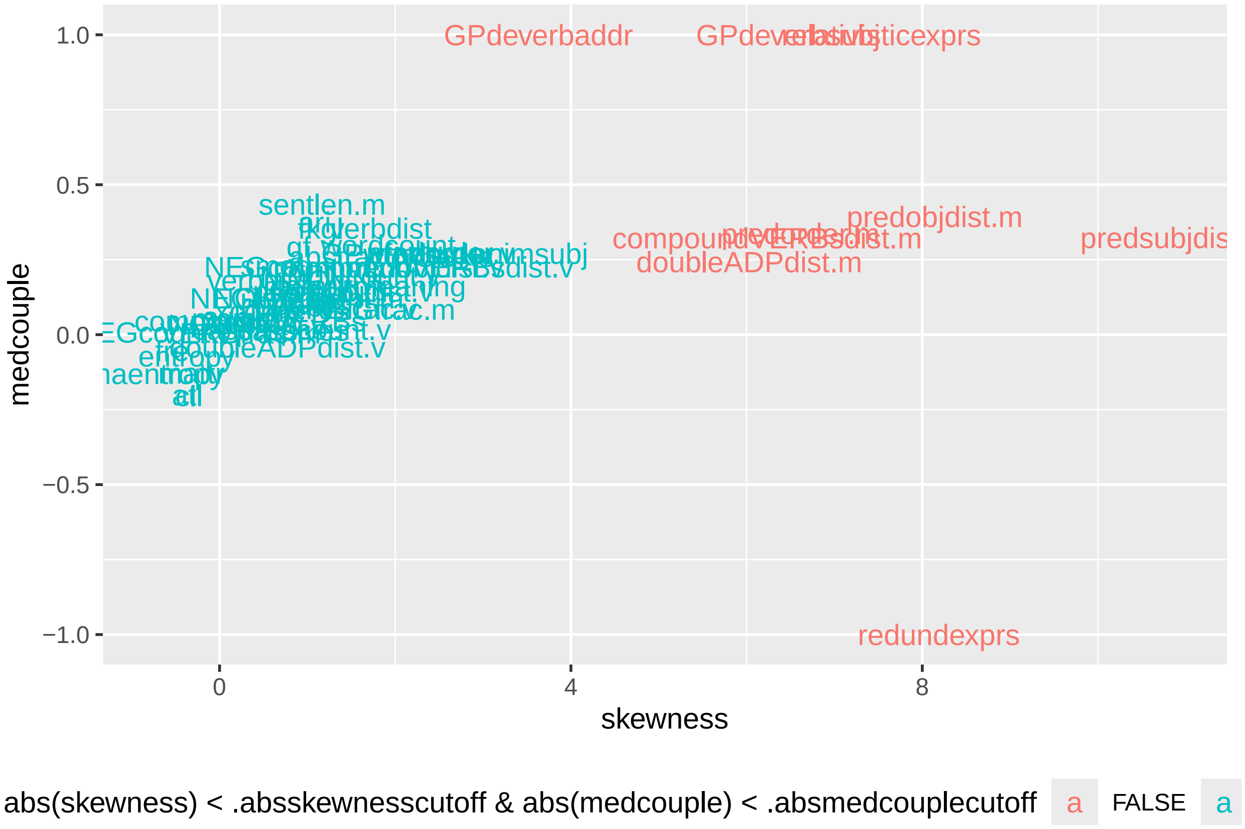
```r
.absskewnesscutoff <- 5
.absmedcouplecutoff <- 0.5

medc <- numeric()
sk <- numeric()
for (i in seq_along(colnames(data_purish))) {
  d <- as.vector(data_purish[i])[[1]]
  medc <- c(medc, mc(d, doScale = TRUE))
  sk <- c(sk, skewness(d))
}

data_skewness <- data.frame(
  feat = colnames(data_purish), medcouple = medc, skewness = sk
)

data_skewness %>%
  ggplot(aes(x = medcouple, y = feat, label = medcouple %>% round(3))) +
  geom_col() +
  geom_label()
```



```
data_skewness %>%
  ggplot(aes(x = skewness, y = feat, label = skewness %>% round(3))) +
  geom_col() +
  geom_label()
```



```
data_skewness %>%
  ggplot(aes(
    x = skewness, y = medcouple, label = feat,
    color = abs(skewness) < .absskewnesscutoff &
      abs(medcouple) < .absmedcouplecutoff
  )) +
  geom_text() +
  theme(legend.position = "bottom")
```



```r
acceptably_skewed <- data_skewness %>%
  filter(abs(skewness) < .absskewnesscutoff &
    abs(medcouple) < .absmedcouplecutoff) %>%
  pull(feat)
acceptably_skewed
```

```
##  [1] "abstractNOUNs"       "caserepcount.v"      "doubleADPdist.v"
##  [4] "GPwordorder"         "xcomp"               "literary"
##  [7] "sentlen.m"           "compoundVERBs"       "compoundVERBsdist.v"
## [10] "passives"            "predorder.v"         "obj"
## [13] "predobjdist.v"       "subj"                "predsubjdist.v"
## [16] "rfpass_animsubj"     "VERBfrac.m"          "VERBfrac.v"
## [19] "NEGcount.m"          "NEGcount.v"          "NEGfrac.m"
## [22] "NEGfrac.v"           "NOUNcount.m"         "NOUNcount.v"
## [25] "NOUNfrac.v"          "verbalNOUNs"         "weakmeaning"
## [28] "activity"            "ari"                 "atl"
## [31] "cli"                 "entropy"             "fkgl"
## [34] "fre"                 "gf"                  "hpoint"
## [37] "maentropy"           "mamr"                "mattr"
## [40] "hapaxes"             "sentcount"           "smog"
## [43] "ttr"                 "verbdist"            "wordcount"
```

```r
data_unskewed <- data_purish %>% select(any_of(acceptably_skewed))

# pairs.panels(data_unskewed, lm = TRUE)
```

Correlations

See Levshina (2015: 353–54).

```r
analyze_correlation <- function(data) {
  cor_matrix <- cor(data)

  cor_tibble_long <- cor_matrix %>%
    as_tibble() %>%
    mutate(feat1 = rownames(cor_matrix)) %>%
    pivot_longer(!feat1, names_to = "feat2", values_to = "cor") %>%
    mutate(abs_cor = abs(cor))

  cor_matrix_upper <- cor_matrix
  cor_matrix_upper[lower.tri(cor_matrix_upper)] <- 0

  cor_tibble_long_upper <- cor_matrix_upper %>%
    as_tibble() %>%
    mutate(feat1 = rownames(cor_matrix)) %>%
    pivot_longer(!feat1, names_to = "feat2", values_to = "cor") %>%
    mutate(abs_cor = abs(cor)) %>%
    filter(feat1 != feat2 & abs_cor > 0)

  list(
    cor_matrix = cor_matrix,
    cor_matrix_upper = cor_matrix_upper,
    cor_tibble_long = cor_tibble_long,
    cor_tibble_long_upper = cor_tibble_long_upper
  )
}


# data_purish <- data_clean %>% select(any_of(selected_features_names)) %>%
#   # remove features expected to have low communalities
#   select(!c(
#     RuleDoubleAdpos.max_allowable_distance,
#     RuleDoubleAdpos.max_allowable_distance.v,
#     RuleGPwordorder,
#     RuleLiteraryStyle,
#     maentropy.v,
#     RuleTooManyNegations.max_negation_frac,
#     RulePredSubjDistance.max_distance,
#     RuleTooManyNegations.max_allowable_negations,
#     RuleTooManyNegations.max_allowable_negations.v,
#     RuleTooManyNominalConstructions.max_allowable_nouns.v,
#     RuleTooFewVerbs.min_verb_frac.v,
#     RulePredObjDistance.max_distance.v,
#     RulePredObjDistance.max_distance,
#     RulePredAtClauseBeginning.max_order.v,
#     RuleInfVerbDistance
#   )) %>%
#   # remove features expected to have low loadings
#   select(!c(
#     RuleMultiPartVerbs.max_distance.v,
#     RulePredSubjDistance.max_distance.v,
#     RuleLongSentences.max_length
#   ))
```

Extremely non-normal data

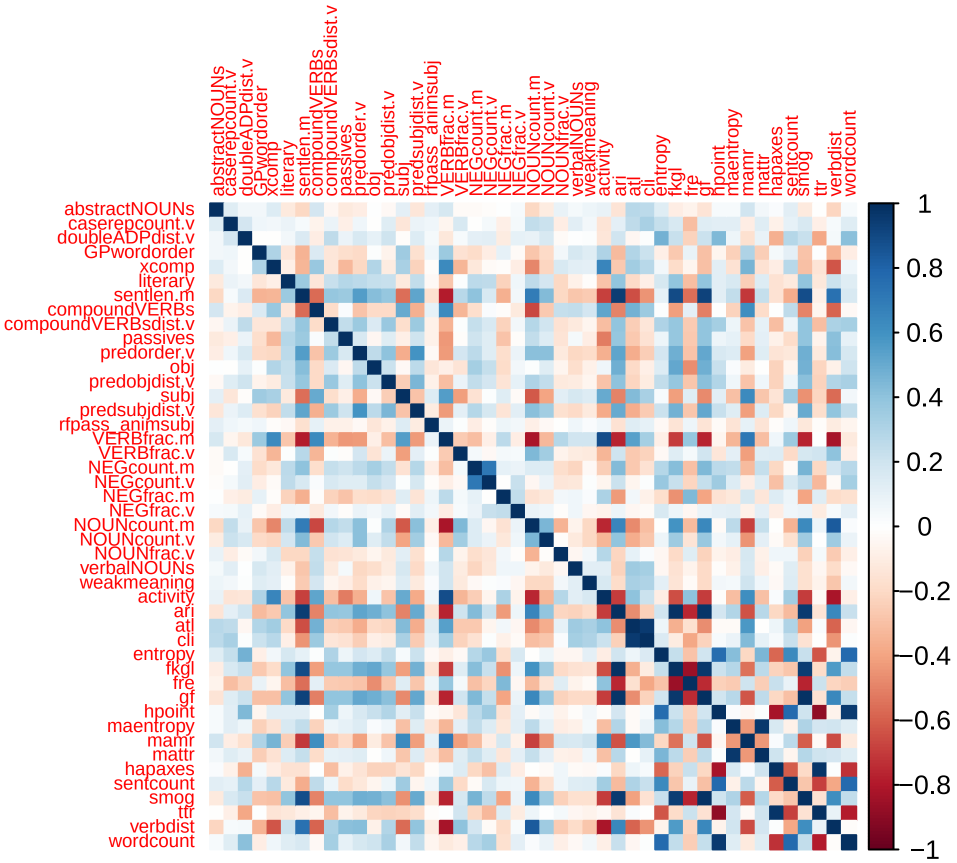
```
# # remove where median == 0?
# keep <- character()
# for (i in seq_along(colnames(data_purish))) {
#   cname <- colnames(data_purish)[i]
#   q <- quantile(data_purish[, i][[1]], probs = 0.10)[[1]]
#   if (q > 0) {
#     keep <- c(keep, cname)
#     cat("keep", cname, "\n")
#   } else {
#     cat("throw out", cname, "\n")
#   }
# }
# data_purish <- data_purish %>% select(any_of(keep))
```

High correlations

```
.hcorrcutoff <- 0.7

library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
corrplot(cor(data_unskewed), method = "color", tl.cex = 0.6)
```



```r
analyze_correlation(data_purish)$cor_tibble_long %>%
  filter(feat1 != feat2 & abs_cor > .hcorrcutoff) %>%
  arrange(feat1, -abs_cor) %>%
  print(n = 100)
```

```
## # A tibble: 82 x 4
##    feat1       feat2          cor abs_cor
##    <chr>       <chr>        <dbl>   <dbl>
##  1 NEGcount.m  NEGcount.v   0.713   0.713
##  2 NEGcount.v  NEGcount.m   0.713   0.713
##  3 NOUNcount.m verbdist     0.831   0.831
##  4 NOUNcount.m VERBfrac.m  -0.820   0.820
##  5 NOUNcount.m activity    -0.758   0.758
##  6 VERBfrac.m  activity     0.889   0.889
##  7 VERBfrac.m  verbdist    -0.822   0.822
##  8 VERBfrac.m  NOUNcount.m -0.820   0.820
##  9 VERBfrac.m  sentlen.m   -0.780   0.780
## 10 VERBfrac.m  ari         -0.765   0.765
## 11 VERBfrac.m  smog        -0.764   0.764
## 12 VERBfrac.m  gf          -0.764   0.764
## 13 VERBfrac.m  mamr         0.710   0.710
## 14 activity    VERBfrac.m   0.889   0.889
## 15 activity    verbdist    -0.818   0.818
## 16 activity    NOUNcount.m -0.758   0.758
## 17 ari         fkgl         0.984   0.984
## 18 ari         gf           0.978   0.978
## 19 ari         smog         0.951   0.951
## 20 ari         sentlen.m    0.944   0.944
## 21 ari         VERBfrac.m  -0.765   0.765
## 22 ari         fre         -0.754   0.754
## 23 atl         cli          0.959   0.959
## 24 cli         atl          0.959   0.959
## 25 entropy     wordcount    0.773   0.773
## 26 entropy     hpoint       0.761   0.761
## 27 fkgl        ari          0.984   0.984
## 28 fkgl        gf           0.968   0.968
## 29 fkgl        smog         0.949   0.949
## 30 fkgl        sentlen.m    0.892   0.892
## 31 fkgl        fre         -0.853   0.853
## 32 fre         fkgl        -0.853   0.853
## 33 fre         smog        -0.763   0.763
## 34 fre         ari         -0.754   0.754
## 35 fre         gf          -0.748   0.748
## 36 gf          smog         0.987   0.987
## 37 gf          ari          0.978   0.978
## 38 gf          fkgl         0.968   0.968
## 39 gf          sentlen.m    0.923   0.923
## 40 gf          VERBfrac.m  -0.764   0.764
## 41 gf          fre         -0.748   0.748
## 42 hapaxes     ttr          0.979   0.979
## 43 hapaxes     hpoint      -0.824   0.824
## 44 hapaxes     wordcount   -0.724   0.724
## 45 hpoint      wordcount    0.958   0.958
## 46 hpoint      ttr         -0.887   0.887
## 47 hpoint      hapaxes     -0.824   0.824
## 48 hpoint      sentcount    0.779   0.779
## 49 hpoint      entropy      0.761   0.761
## 50 maentropy   mattr        0.964   0.964
## 51 mamr        VERBfrac.m   0.710   0.710
## 52 mamr        sentlen.m   -0.702   0.702
## 53 mattr       maentropy    0.964   0.964
## 54 predorder.m sentlen.m    0.714   0.714
## 55 sentcount   hpoint       0.779   0.779
## 56 sentcount   wordcount    0.779   0.779
## 57 sentlen.m   ari          0.944   0.944
## 58 sentlen.m   gf           0.923   0.923
## 59 sentlen.m   fkgl         0.892   0.892
## 60 sentlen.m   smog         0.881   0.881
## 61 sentlen.m   VERBfrac.m  -0.780   0.780
## 62 sentlen.m   verbdist     0.745   0.745
## 63 sentlen.m   predorder.m  0.714   0.714
## 64 sentlen.m   mamr        -0.702   0.702
## 65 smog        gf           0.987   0.987
## 66 smog        ari          0.951   0.951
## 67 smog        fkgl         0.949   0.949
## 68 smog        sentlen.m    0.881   0.881
## 69 smog        VERBfrac.m  -0.764   0.764
## 70 smog        fre         -0.763   0.763
## 71 ttr         hapaxes      0.979   0.979
## 72 ttr         hpoint      -0.887   0.887
## 73 ttr         wordcount   -0.783   0.783
## 74 verbdist    NOUNcount.m  0.831   0.831
## 75 verbdist    VERBfrac.m  -0.822   0.822
## 76 verbdist    activity    -0.818   0.818
## 77 verbdist    sentlen.m    0.745   0.745
## 78 wordcount   hpoint       0.958   0.958
## 79 wordcount   ttr         -0.783   0.783
## 80 wordcount   sentcount    0.779   0.779
## 81 wordcount   entropy      0.773   0.773
## 82 wordcount   hapaxes     -0.724   0.724
```

exclude:

- **ari:** corr. w/ RuleLongSentences.max_length > 0.94; sentence length seems more universal, let's make it a substitute
- **gf:** corr. w/ RuleLongSentences.max_length > 0.92; sentence length seems more universal, let's make it a substitute
- **maentropy:** corr. w/ mattr > 0.96, but mattr is implemented in QuitaUp. besides, the interesting thing about maentropy is its variation
- **smog:** corr. w/ fkgl almost 0.95, but fkgl coefficients adjusted for Czech are available
- **atl:** corr. w/ cli around 0.96; unlike cli, atl is not a readability metric

```r
high_correlations <- findCorrelation(
  cor(data_purish),
  verbose = TRUE, cutoff = .hcorrcutoff
)
```

```
## Compare row 10  and column  38 with corr  0.944
##   Means:  0.384 vs 0.196 so flagging column 10
## Compare row 38  and column  44 with corr  0.978
##   Means:  0.37 vs 0.189 so flagging column 38
## Compare row 44  and column  51 with corr  0.987
##   Means:  0.357 vs 0.182 so flagging column 44
## Compare row 51  and column  42 with corr  0.949
##   Means:  0.338 vs 0.176 so flagging column 51
## Compare row 42  and column  43 with corr  0.853
##   Means:  0.31 vs 0.171 so flagging column 42
## Compare row 26  and column  53 with corr  0.822
##   Means:  0.304 vs 0.164 so flagging column 26
## Compare row 53  and column  37 with corr  0.818
##   Means:  0.283 vs 0.158 so flagging column 53
## Compare row 37  and column  32 with corr  0.758
##   Means:  0.258 vs 0.153 so flagging column 37
## Compare row 39  and column  40 with corr  0.959
##   Means:  0.227 vs 0.15 so flagging column 39
## Compare row 50  and column  54 with corr  0.779
##   Means:  0.197 vs 0.146 so flagging column 50
## Compare row 28  and column  29 with corr  0.713
##   Means:  0.175 vs 0.145 so flagging column 28
## Compare row 41  and column  54 with corr  0.773
##   Means:  0.198 vs 0.143 so flagging column 41
## Compare row 54  and column  45 with corr  0.958
##   Means:  0.169 vs 0.14 so flagging column 54
## Compare row 45  and column  52 with corr  0.887
##   Means:  0.148 vs 0.139 so flagging column 45
## Compare row 52  and column  49 with corr  0.979
##   Means:  0.136 vs 0.139 so flagging column 49
## Compare row 46  and column  48 with corr  0.964
##   Means:  0.149 vs 0.139 so flagging column 46
## All correlations <= 0.7
```

```
names(data_purish)[high_correlations]
```

```
##  [1] "sentlen.m"  "ari"        "gf"         "smog"       "fkgl"
##  [6] "VERBfrac.m" "verbdist"   "activity"   "atl"        "sentcount"
## [11] "NEGcount.m" "entropy"    "wordcount"  "hpoint"     "hapaxes"
## [16] "maentropy"
```

```
data_pureish_striphigh <- data_purish %>% select(!all_of(high_correlations))

analyze_correlation(data_pureish_striphigh)$cor_tibble_long %>%
  filter(feat1 != feat2 & abs_cor > .hcorrcutoff) %>%
  arrange(feat1, -abs_cor) %>%
  print(n = 100)
```

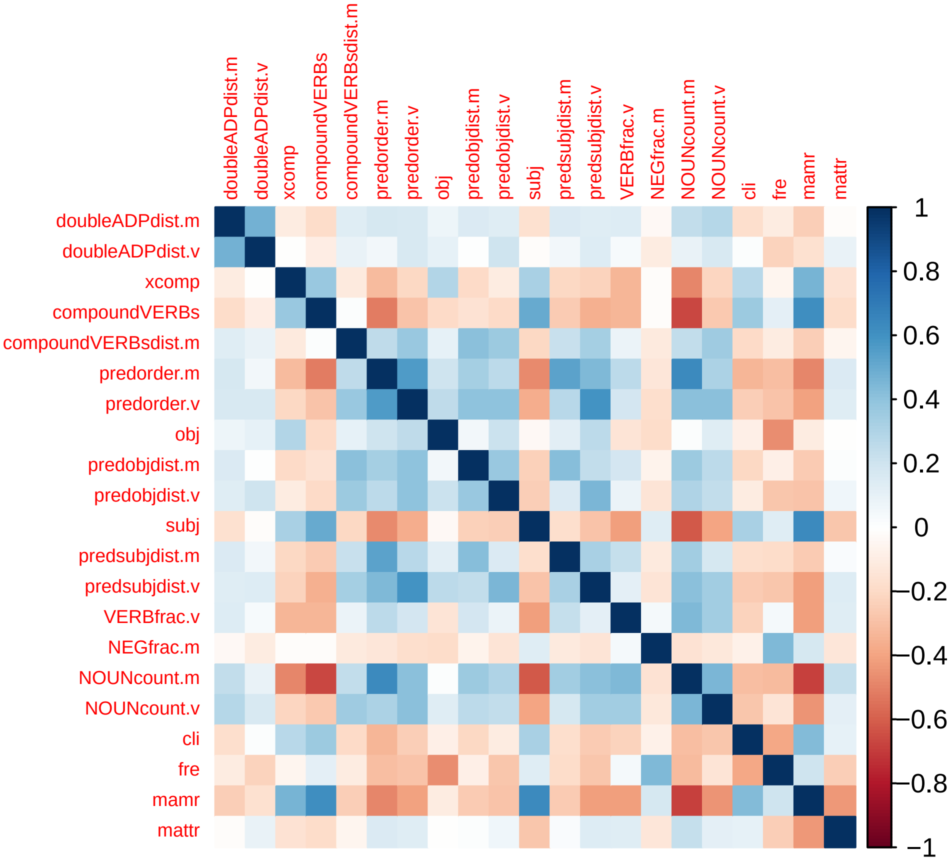
```
## # A tibble: 0 x 4
## # i 4 variables: feat1 <chr>, feat2 <chr>, cor <dbl>, abs_cor <dbl>
```

Low correlations

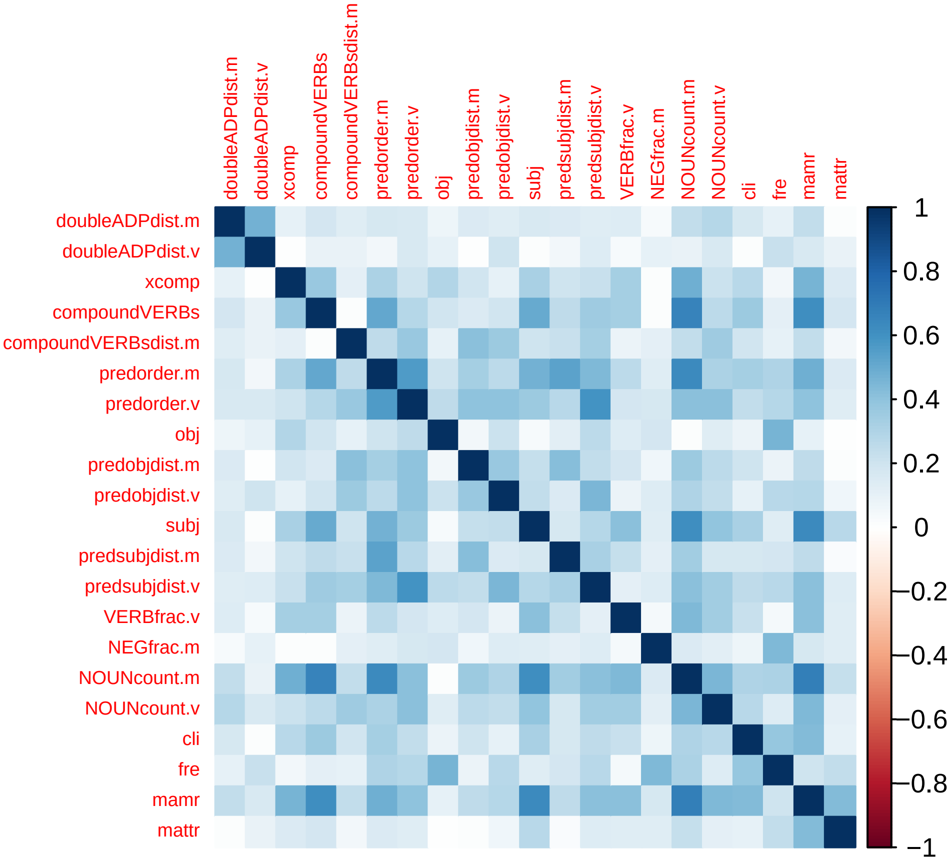
```
.lcorrcutoff <- 0.4

low_correlating_features <- analyze_correlation(data_pureish_striphigh)$
  cor_tibble_long %>%
  filter(feat1 != feat2) %>%
  group_by(feat1) %>%
  summarize(max_cor = max(abs_cor)) %>%
  filter(max_cor < .lcorrcutoff) %>%
  pull(feat1)

feature_importances %>% filter(feat_name %in% low_correlating_features)
```

```
## # A tibble: 1 x 2
##   feat_name p_value
##   <chr>       <dbl>
## 1 ttr        0.0461
```

```
data_pure <- data_pureish_striphigh %>%
  select(!any_of(low_correlating_features))

colnames(data_pure) <- prettify_feat_name_vector(colnames(data_pure))

corrplot(cor(data_pure), method = "color", tl.cex = 0.6)
```



```
corrplot(cor(data_pure) %>% abs(), method = "color", tl.cex = 0.6)
```



Verbalness FA

```
# data_verbalness <- data_pure %>%
#   select(VERBfrac.m, NOUNcount.m, activity, verbdist) %>%
#   # mutate(neg.NOUNcount.m = -NOUNcount.m, neg.verbdist = -verbdist) %>%
#   # select(!c(NOUNcount.m, verbdist)) %>%
#   mutate(across(VERBfrac.m:verbdist, ~ scale(.x)[, 1]))

# corrplot(cor(data_verbalness), method = "color", tl.cex = 0.6)

# data_verbalness %>%
#   cor() %>%
#   det()
# KMO(data_verbalness)

# mult.norm(data_verbalness %>% as.data.frame())$mult.test

# fa.parallel(data_verbalness, fm = "pa", fa = "fa", n.iter = 20)

# fa_verbalness <- fa(
#   data_verbalness,
#   nfactors = 2,
#   fm = "pa",
#   rotate = "promax",
#   oblique.scores = TRUE,
#   scores = "tenBerge",
#   n.iter = 100
# )
# fa_verbalness

# data_verbalness_parcels <- data_verbalness %>%
#   rowid_to_column("ID") %>%
#   pivot_longer(!ID, names_to = "name", values_to = "value") %>%
#   group_by(ID) %>%
#   summarize(
#     verbalness = weighted.mean(value, fa_verbalness$loadings[, "PA1"]),
#     nominalness = weighted.mean(value, fa_verbalness$loadings[, "PA2"])
#   ) %>%
#   ungroup() %>%
#   select(!ID)

# data_pure <- data_pure %>%
#   bind_cols(data_verbalness_parcels) %>%
#   select(!c(VERBfrac.m, NOUNcount.m, activity, verbdist))
```

Correlation matrix determinant

determinants > 0.00001 = 1e-5 generally indicate multicollinearity is probably not a problem (Watkins 2021: 61).

```
data_pure %>%
  cor() %>%
  det() %>%
  print(digits = 5)
```

```
## [1] 0.00010501
```

Bartlett's test of sphericity

null hypothesis: the correlation matrix is an identity matrix, i.e. random. sensitive to sample size, should be considered a minimal standard (Watkins 2021: 61).

```
cortest.bartlett(data_pure)
```

```
## R was not square, finding R from data
```

```
## $chisq
## [1] 6817.642
##
## $p.value
## [1] 0
##
## $df
## [1] 210
```

Kaiser-Meyer-Olkin measure of sampling adequacy

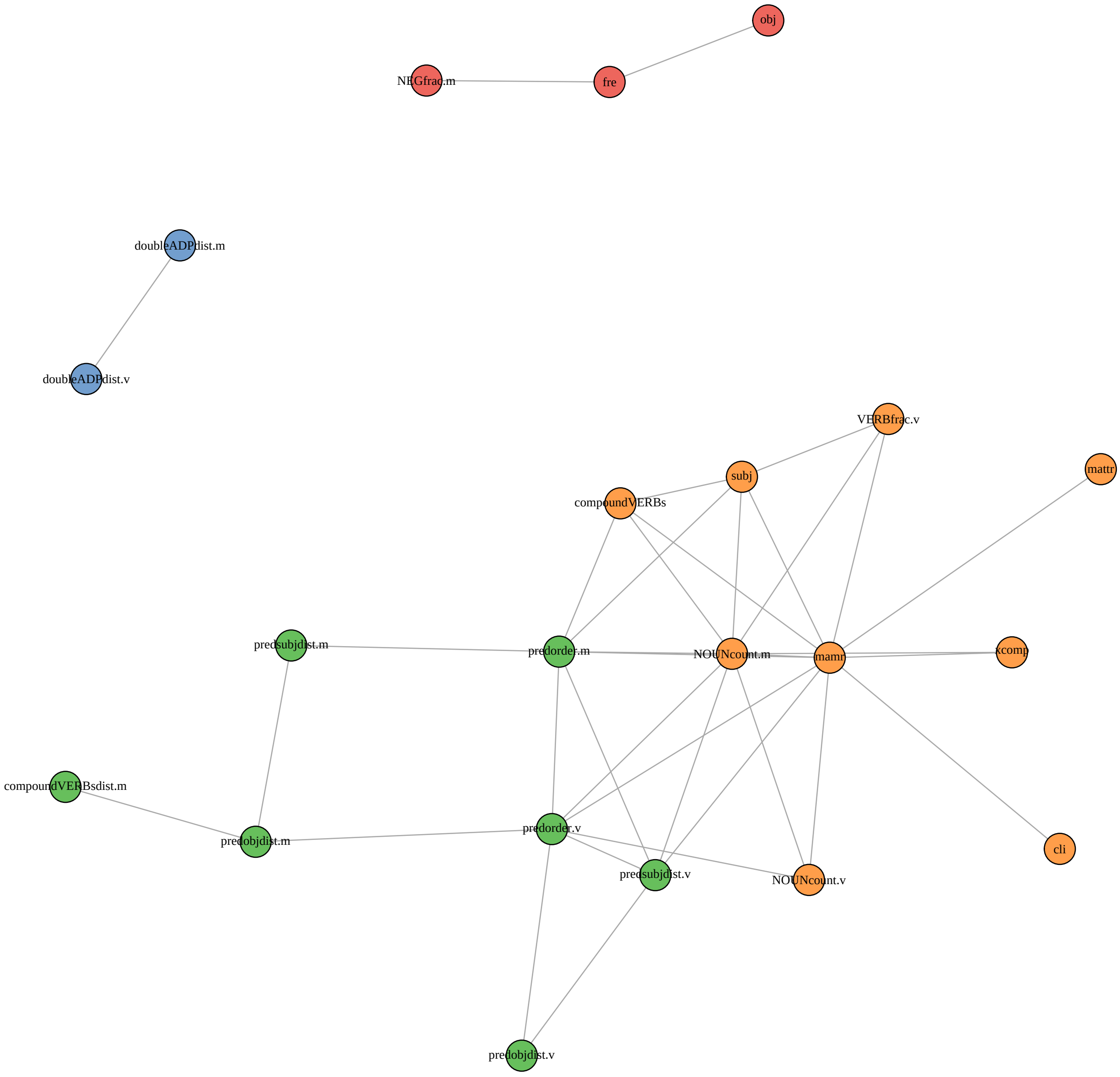
there are debates about which values KMO values are acceptable. Kaiser (1974) suggested that KMO < .50 are unacceptable but other measurement specialists recomment a minimum value of .60 for acceptability with values >= .70 preferred (Watkins 2021: 61).

```
KMO(data_pure)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = data_pure)
## Overall MSA =  0.79
## MSA for each item = 
##     doubleADPdist.m     doubleADPdist.v               xcomp       compoundVERBs 
##                0.73                0.60                0.87                0.80 
## compoundVERBsdist.m         predorder.m         predorder.v                 obj 
##                0.82                0.83                0.84                0.46 
##       predobjdist.m       predobjdist.v                subj      predsubjdist.m 
##                0.77                0.86                0.92                0.73 
##      predsubjdist.v          VERBfrac.v           NEGfrac.m         NOUNcount.m 
##                0.85                0.88                0.73                0.84 
##         NOUNcount.v                 cli                 fre                mamr 
##                0.91                0.60                0.54                0.85 
##               mattr 
##                0.64
```

Visualisation

```
my_colors <- paletteer::paletteer_d("ggthemes::Classic_10_Medium")

network_edges <- analyze_correlation(data_pure)$cor_tibble_long_upper %>%
  filter(abs_cor > .lcorrcutoff)

network <- graph_from_data_frame(
  network_edges,
  directed = FALSE
)
E(network)$weight <- network_edges$abs_cor
network_communities <- cluster_optimal(network)

network_membership <- membership(network_communities)

plot(
  network,
  layout = layout.fruchterman.reingold,
  vertex.color = map(
    network_communities$membership,
    function(x) my_colors[x]
  ) %>% unlist(use.names = FALSE),
  vertex.size = 6,
  vertex.label.color = "black",
  vertex.label.cex = 0.7
)
```



Scaling

```
data_scaled <- data_pure %>%
  mutate(across(seq_along(data_pure), ~ scale(.x)[, 1]))

final_collist <- data_scaled %>% colnames()
```

Check for normality

```
mult.norm(data_scaled %>% as.data.frame())$mult.test
```

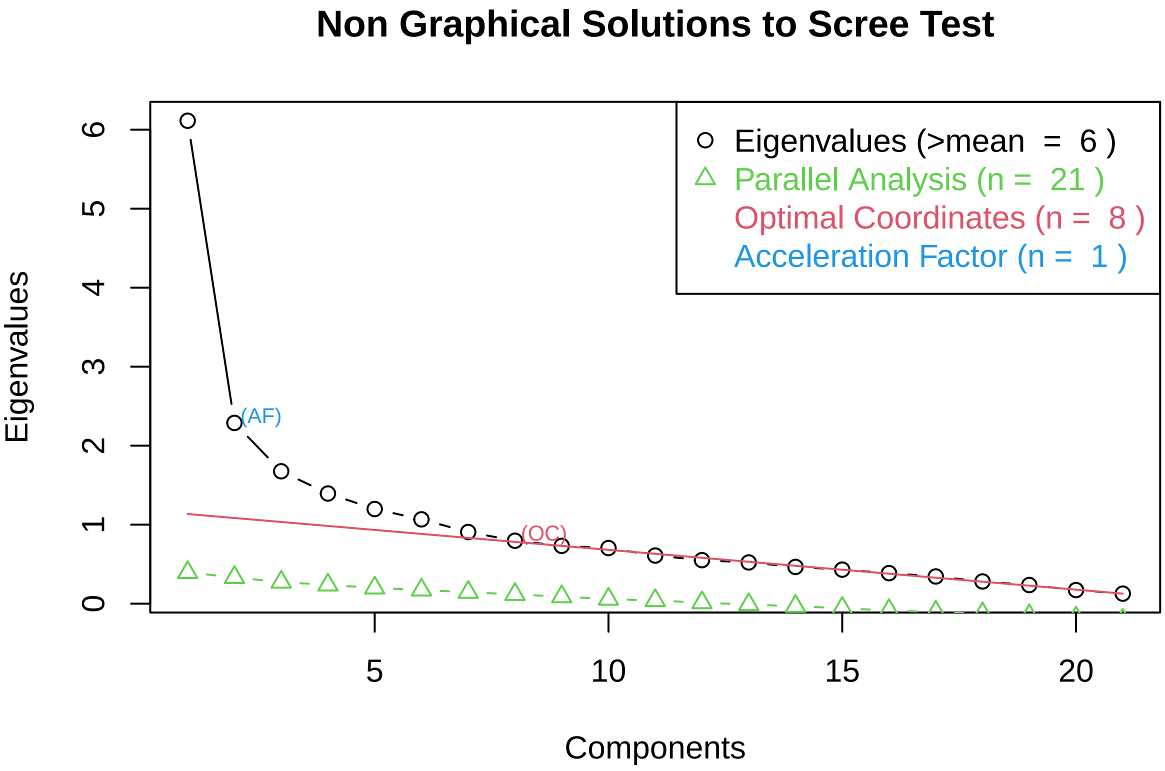
```
##           Beta-hat      kappa p-val
## Skewness  706.5797 88675.7571     0
## Kurtosis 1579.0911   483.8665     0
```

Low (null) p-values show that we can reject the hypothesis that the data would be in a multivariate normal distribution. I.e. the distribution isn't multivariate normal.

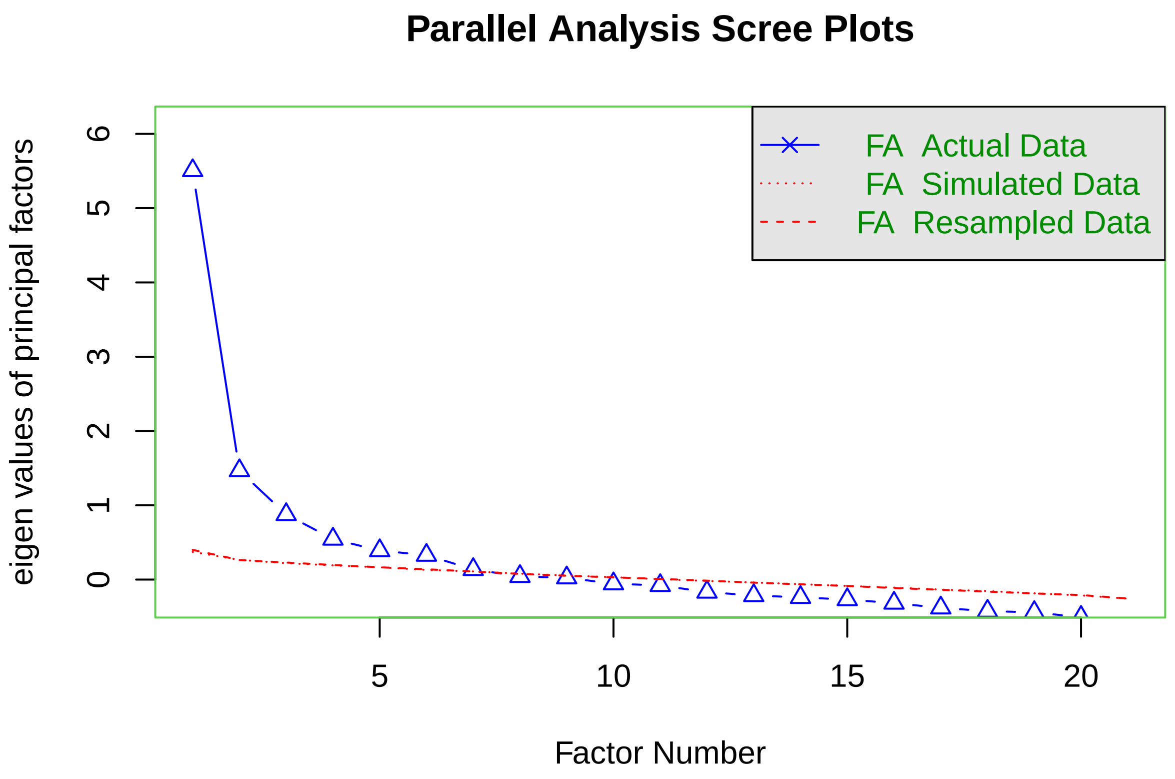
FA

No. of factors

```
eigen <- eigen(cor(data_scaled))
par <- nFactors::parallel(
  subject = nrow(data_scaled),
  var = ncol(data_scaled),
  rep = 100,
  quantile = .95,
  model = "factors"
)
scree <- nScree(x = eigen$values, aparallel = par$eigen$qevpea)
plotnScree(scree)
```



```
fa.parallel(data_scaled, fm = "pa", fa = "fa", n.iter = 20)
```



```
## Parallel analysis suggests that the number of factors =  7  and the number of components =  NA
```

Model

https://www.rdocumentation.org/packages/psych/versions/2.5.3/topics/fa

```
# appears to be the happiest when nfactors = 6 or 7
# throws the The estimated weights for the factor scores are probably incorrect.
# Try a different factor score estimation method. warning otherwise
fa_res <- fa(
  data_scaled,
  nfactors = 7,
  fm = "pa",
  rotate = "promax",
  oblique.scores = TRUE,
  scores = "tenBerge",
  n.iter = 100
)
```

```
## Loading required namespace: GPArotation
```

```
fa_res
```

```
## Factor Analysis with confidence intervals using method = fa(r = data_scaled, nfactors = 7, n.iter =
##     scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Factor Analysis using method =  pa
## Call: fa(r = data_scaled, nfactors = 7, n.iter = 100, rotate = "promax",
##     scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##                        PA1   PA2   PA7   PA3   PA6   PA5   PA4   h2    u2 com
## doubleADPdist.m      -0.11 -0.03 -0.14  0.04  0.09  0.03  0.68 0.52 0.478 1.2
## doubleADPdist.v       0.12  0.07  0.16 -0.07 -0.07 -0.06  0.76 0.58 0.424 1.2
## xcomp                0.46  0.13 -0.14  0.02 -0.11  0.36  0.00 0.43 0.575 2.4
## compoundVERBs        0.68  0.12 -0.13  0.40 -0.18 -0.20 -0.05 0.68 0.316 2.1
## compoundVERBsdist.m  0.07  0.00  0.14  0.70 -0.01 -0.04 -0.03 0.51 0.495 1.1
## predorder.m         -0.31  0.08  0.18 -0.04  0.52  0.10 -0.06 0.68 0.324 2.1
## predorder.v         -0.05  0.01  0.50  0.30  0.07  0.03 -0.01 0.57 0.433 1.7
## obj                  0.07  0.19  0.04 -0.08  0.03  0.90 -0.03 0.86 0.143 1.1
## predobjdist.m        0.00  0.03  0.03  0.49  0.34 -0.09 -0.05 0.45 0.551 1.9
## predobjdist.v        0.06  0.07  0.40  0.32 -0.02  0.00  0.05 0.36 0.635 2.1
## subj                 0.81 -0.06  0.09 -0.12  0.08 -0.07  0.11 0.58 0.422 1.2
## predsubjdist.m       0.00  0.10  0.00  0.06  0.67  0.00  0.03 0.48 0.519 1.1
## predsubjdist.v       0.04 -0.09  0.78  0.12  0.03  0.01  0.00 0.65 0.354 1.1
## VERBfrac.v          -0.59 -0.03 -0.22  0.08  0.09 -0.12  0.03 0.35 0.653 1.5
## NEGfrac.m            0.07 -0.40  0.02 -0.12  0.00 -0.10  0.02 0.20 0.797 1.4
## NOUNcount.m         -0.73  0.16 -0.01  0.01  0.26 -0.06  0.01 0.77 0.230 1.4
## NOUNcount.v         -0.42  0.02 -0.03  0.37 -0.11  0.07  0.11 0.42 0.580 2.4
## cli                  0.39  0.60 -0.11 -0.11  0.05 -0.22  0.02 0.59 0.408 2.2
## fre                  0.02 -0.95  0.03  0.07 -0.25 -0.21 -0.07 0.99 0.011 1.3
## mamr                 0.87 -0.03 -0.09 -0.02  0.16 -0.05 -0.04 0.79 0.206 1.1
## mattr               -0.42  0.28  0.09 -0.13 -0.18 -0.10 -0.04 0.28 0.716 2.7
##
##                        PA1  PA2  PA7  PA3  PA6  PA5  PA4
## SS loadings           3.74 1.59 1.46 1.43 1.27 1.15 1.08
## Proportion Var        0.18 0.08 0.07 0.07 0.06 0.05 0.05
## Cumulative Var        0.18 0.25 0.32 0.39 0.45 0.51 0.56
## Proportion Explained  0.32 0.14 0.12 0.12 0.11 0.10 0.09
## Cumulative Proportion 0.32 0.45 0.58 0.70 0.81 0.91 1.00
##
##  With factor correlations of
##       PA1   PA2   PA7   PA3   PA6   PA5   PA4
## PA1  1.00 -0.04 -0.54 -0.34 -0.43 -0.09 -0.27
## PA2 -0.04  1.00  0.31  0.02 -0.13  0.09  0.08
## PA7 -0.54  0.31  1.00  0.37  0.37  0.34  0.23
## PA3 -0.34  0.02  0.37  1.00  0.37  0.26  0.30
## PA6 -0.43 -0.13  0.37  0.37  1.00  0.16  0.13
## PA5 -0.09  0.09  0.34  0.26  0.16  1.00  0.21
## PA4 -0.27  0.08  0.23  0.30  0.13  0.21  1.00
##
## Mean item complexity =  1.6
## Test of the hypothesis that 7 factors are sufficient.
##
## df null model =  210  with the objective function =  9.16 with Chi Square =  6817.64
## df of  the model are 84  and the objective function was  0.85
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.04
##
## The harmonic n.obs is  753 with the empirical chi square  212.23  with prob <  4.5e-13
## The total n.obs was  753  with Likelihood Chi Square =  631.99  with prob <  6.1e-85
##
## Tucker Lewis Index of factoring reliability =  0.791
## RMSEA index =  0.093  and the 90 % confidence intervals are  0.086 0.1
## BIC =  75.57
## Fit based upon off diagonal values = 0.99
## Measures of factor score adequacy
##                                                    PA1  PA2  PA7  PA3  PA6  PA5
## Correlation of (regression) scores with factors   0.96 0.99 0.89 0.86 0.86 0.93
## Multiple R square of scores with factors          0.93 0.97 0.80 0.74 0.74 0.86
## Minimum correlation of possible factor scores     0.85 0.95 0.59 0.49 0.48 0.71
##                                                    PA4
## Correlation of (regression) scores with factors   0.85
## Multiple R square of scores with factors          0.72
## Minimum correlation of possible factor scores     0.44
##
##  Coefficients and bootstrapped confidence intervals
##                       low   PA1 upper   low   PA2 upper   low   PA7 upper   low
## doubleADPdist.m     -0.17 -0.11 -0.02 -0.10 -0.03  0.01 -0.37 -0.14  0.16 -0.13
## doubleADPdist.v     -0.01  0.12  0.18  0.03  0.07  0.16 -0.08  0.16  0.36 -0.21
## xcomp                0.26  0.46  0.72  0.03  0.13  0.20 -0.39 -0.14  0.18 -0.16
## compoundVERBs        0.47  0.68  0.94  0.03  0.12  0.18 -0.42 -0.13  0.13  0.02
## compoundVERBsdist.m -0.02  0.07  0.20 -0.09  0.00  0.07 -0.37  0.14  1.07  0.14
## predorder.m         -0.62 -0.31 -0.11 -0.06  0.08  0.16 -0.21  0.18  0.92 -0.14
## predorder.v         -0.25 -0.05  0.08 -0.07  0.01  0.12 -0.11  0.50  1.52 -0.09
## obj                 -0.04  0.07  0.27  0.13  0.19  0.24 -0.47  0.04  0.99 -0.37
## predobjdist.m       -0.10  0.00  0.13 -0.09  0.03  0.12 -0.42  0.03  0.88  0.08
## predobjdist.v       -0.05  0.06  0.16  0.00  0.07  0.18 -0.01  0.40  1.21 -0.03
## subj                 0.53  0.81  0.97 -0.11 -0.06  0.02 -0.42  0.09  0.42 -0.23
## predsubjdist.m      -0.14  0.00  0.08 -0.03  0.10  0.18 -0.39  0.00  0.63 -0.11
## predsubjdist.v      -0.22  0.04  0.18 -0.13 -0.09  0.09  0.17  0.78  1.59 -0.19
## VERBfrac.v          -0.76 -0.59 -0.39 -0.13 -0.03  0.04 -0.52 -0.22  0.09 -0.07
## NEGfrac.m           -0.06  0.07  0.15 -0.49 -0.40 -0.31 -0.40  0.02  0.31 -0.34
## NOUNcount.m         -0.99 -0.73 -0.46  0.08  0.16  0.22 -0.21 -0.01  0.34 -0.13
## NOUNcount.v         -0.55 -0.42 -0.24 -0.07  0.02  0.09 -0.51 -0.03  0.67  0.12
## cli                  0.27  0.39  0.47  0.52  0.60  0.71 -0.86 -0.11  0.35 -0.31
## fre                 -0.05  0.02  0.07 -1.04 -0.95 -0.80 -0.30  0.03  0.19 -0.08
## mamr                 0.60  0.87  1.11 -0.11 -0.03  0.03 -0.61 -0.09  0.32 -0.12
## mattr               -0.63 -0.42 -0.24  0.21  0.28  0.40 -0.31  0.09  0.37 -0.31
##                       PA3 upper   low   PA6 upper   low   PA5 upper   low   PA4
## doubleADPdist.m      0.04  0.20 -0.08  0.09  0.18 -0.07  0.03  0.09  0.49  0.68
## doubleADPdist.v     -0.07  0.13 -0.19 -0.07  0.02 -0.14 -0.06  0.08  0.48  0.76
## xcomp                0.02  0.11 -0.37 -0.11  0.05  0.13  0.36  0.72 -0.09  0.00
## compoundVERBs        0.40  1.00 -0.56 -0.18  0.06 -0.61 -0.20  0.06 -0.12 -0.05
## compoundVERBsdist.m  0.70  1.57 -0.21 -0.01  0.29 -0.24 -0.04  0.09 -0.12 -0.03
## predorder.m         -0.04  0.14  0.03  0.52  1.19 -0.12  0.10  0.43 -0.19 -0.06
## predorder.v          0.30  0.96 -0.29  0.07  0.61 -0.21  0.03  0.40 -0.11 -0.01
## obj                 -0.08  0.07 -0.15  0.03  0.27  0.29  0.90  1.84 -0.08 -0.03
## predobjdist.m        0.49  1.21 -0.09  0.34  0.88 -0.34 -0.09  0.07 -0.15 -0.05
## predobjdist.v        0.32  0.91 -0.28 -0.02  0.42 -0.16  0.00  0.22 -0.01  0.05
## subj                -0.12  0.01 -0.17  0.08  0.23 -0.21 -0.07  0.18  0.02  0.11
## predsubjdist.m       0.06  0.35  0.24  0.67  1.27 -0.09  0.00  0.12 -0.07  0.03
## predsubjdist.v       0.12  0.69 -0.37  0.03  0.74 -0.26  0.01  0.48 -0.09  0.00
## VERBfrac.v           0.08  0.22 -0.15  0.09  0.27 -0.61 -0.12  0.16 -0.07  0.03
## NEGfrac.m           -0.12  0.05 -0.18  0.00  0.14 -0.28 -0.10  0.05 -0.09  0.02
## NOUNcount.m          0.01  0.12  0.04  0.26  0.62 -0.26 -0.06  0.07 -0.06  0.01
## NOUNcount.v          0.37  0.78 -0.38 -0.11  0.18 -0.22  0.07  0.26  0.01  0.11
## cli                 -0.11  0.04 -0.11  0.05  0.20 -0.56 -0.22  0.00 -0.03  0.02
## fre                  0.07  0.26 -0.53 -0.25 -0.08 -0.62 -0.21  0.06 -0.13 -0.07
## mamr                -0.02  0.10 -0.19  0.16  0.43 -0.17 -0.05  0.05 -0.13 -0.04
## mattr               -0.13  0.00 -0.49 -0.18  0.12 -0.31 -0.10  0.10 -0.13 -0.04
##                      upper
## doubleADPdist.m      0.97
## doubleADPdist.v      1.05
## xcomp                0.09
## compoundVERBs        0.03
## compoundVERBsdist.m  0.08
## predorder.m          0.02
## predorder.v          0.09
## obj                  0.06
## predobjdist.m        0.03
## predobjdist.v        0.13
## subj                 0.18
## predsubjdist.m       0.11
## predsubjdist.v       0.10
## VERBfrac.v           0.15
## NEGfrac.m            0.10
## NOUNcount.m          0.07
## NOUNcount.v          0.24
## cli                  0.09
## fre                 -0.02
## mamr                 0.02
## mattr                0.07
##
##  Interfactor correlations and bootstrapped confidence intervals
##         lower estimate upper
## PA1-PA2 -0.86   -0.040 0.029
## PA1-PA7 -0.74   -0.542 0.307
## PA1-PA3 -0.64   -0.340 0.228
## PA1-PA6 -0.60   -0.429 0.276
## PA1-PA5 -0.48   -0.093 0.149
## PA1-PA4 -0.56   -0.268 0.401
## PA2-PA7 -0.32    0.307 0.791
## PA2-PA3 -0.28    0.017 0.685
## PA2-PA6 -0.22   -0.130 0.668
## PA2-PA5 -0.13    0.085 0.563
## PA2-PA4 -0.42    0.083 0.568
## PA7-PA3 -0.35    0.366 0.415
## PA7-PA6 -0.30    0.365 0.520
## PA7-PA5 -0.25    0.337 0.466
## PA7-PA4 -0.40    0.233 0.496
## PA3-PA6 -0.27    0.366 0.486
## PA3-PA5 -0.23    0.257 0.434
## PA3-PA4 -0.38    0.297 0.413
## PA6-PA5 -0.19    0.165 0.417
## PA6-PA4 -0.27    0.132 0.372
## PA5-PA4 -0.27    0.210 0.388
```

Loadings

```
fa_res$loadings
```

```
##
## Loadings:
##                      PA1    PA2    PA7    PA3    PA6    PA5    PA4
## doubleADPdist.m     -0.105        -0.141                       0.681
## doubleADPdist.v      0.120         0.161                       0.762
## xcomp                0.458  0.125 -0.143        -0.114  0.355
## compoundVERBs        0.684  0.120 -0.128  0.396 -0.176 -0.196
## compoundVERBsdist.m                0.137  0.700
## predorder.m         -0.311         0.184         0.521  0.100
## predorder.v                        0.502  0.300
## obj                         0.186                       0.897
## predobjdist.m                             0.489  0.336
## predobjdist.v                      0.404  0.324
## subj                 0.806               -0.125                0.112
## predsubjdist.m              0.101                0.669
## predsubjdist.v                     0.782  0.116
## VERBfrac.v          -0.590        -0.220               -0.124
## NEGfrac.m                  -0.403        -0.120
## NOUNcount.m         -0.727  0.160                0.256
## NOUNcount.v         -0.419                0.371 -0.111         0.109
## cli                  0.386  0.604 -0.113 -0.112        -0.217
## fre                        -0.948               -0.254 -0.213
## mamr                 0.869                       0.157
## mattr               -0.421  0.276        -0.130 -0.177
##
##                  PA1   PA2   PA7   PA3   PA6   PA5   PA4
## SS loadings    3.606 1.635 1.253 1.324 1.112 1.137 1.092
## Proportion Var 0.172 0.078 0.060 0.063 0.053 0.054 0.052
## Cumulative Var 0.172 0.250 0.309 0.372 0.425 0.479 0.531
```

```
for (i in 1:fa_res$factors) {
  cat("\n-----", colnames(fa_res$loadings)[i], "-----\n")

  loadings <- fa_res$loadings[, i]
  load_df <- data.frame(loading = loadings)

  load_df_filtered <- load_df %>%
    mutate(abs_l = abs(loading)) %>%
    mutate(str = case_when(
      abs_l > 0.7 ~ "***",
      abs_l <= 0.7 & abs_l > 0.5 ~ "** ",
      abs_l <= 0.5 & abs_l > 0.3 ~ "*  ",
      abs_l <= 0.3 & abs_l > 0.1 ~ ".  ",
      .default = ""
    )) %>%
    arrange(-abs_l) %>%
    filter(abs_l > 0.1)

  load_df_filtered %>%
    mutate(across(c(loading, abs_l), ~ round(.x, 3))) %>%
    print()

  cat("\n")
}
```

```
##
## ----- PA1 -----
##                  loading abs_l str
## mamr              0.869 0.869 ***
## subj              0.806 0.806 ***
## NOUNcount.m      -0.727 0.727 ***
## compoundVERBs     0.684 0.684 **
## VERBfrac.v       -0.590 0.590 **
## xcomp             0.458 0.458 *
## mattr            -0.421 0.421 *
## NOUNcount.v      -0.419 0.419 *
## cli               0.386 0.386 *
## predorder.m      -0.311 0.311 *
## doubleADPdist.v   0.120 0.120 .
## doubleADPdist.m  -0.105 0.105 .
##
##
## ----- PA2 -----
##                loading abs_l str
## fre             -0.948 0.948 ***
## cli              0.604 0.604 **
## NEGfrac.m       -0.403 0.403 *
## mattr            0.276 0.276 .
## obj              0.186 0.186 .
## NOUNcount.m      0.160 0.160 .
## xcomp            0.125 0.125 .
## compoundVERBs    0.120 0.120 .
## predsubjdist.m   0.101 0.101 .
##
##
## ----- PA7 -----
##                     loading abs_l str
## predsubjdist.v        0.782 0.782 ***
## predorder.v           0.502 0.502 **
## predobjdist.v         0.404 0.404 *
## VERBfrac.v           -0.220 0.220 .
## predorder.m           0.184 0.184 .
## doubleADPdist.v       0.161 0.161 .
## xcomp                -0.143 0.143 .
## doubleADPdist.m      -0.141 0.141 .
## compoundVERBsdist.m   0.137 0.137 .
## compoundVERBs        -0.128 0.128 .
## cli                  -0.113 0.113 .
##
##
## ----- PA3 -----
##                     loading abs_l str
## compoundVERBsdist.m   0.700 0.700 ***
## predobjdist.m         0.489 0.489 *
## compoundVERBs         0.396 0.396 *
## NOUNcount.v           0.371 0.371 *
## predobjdist.v         0.324 0.324 *
## predorder.v           0.300 0.300 *
## mattr                -0.130 0.130 .
## subj                 -0.125 0.125 .
## NEGfrac.m            -0.120 0.120 .
## predsubjdist.v        0.116 0.116 .
## cli                  -0.112 0.112 .
##
##
## ----- PA6 -----
##                loading abs_l str
## predsubjdist.m   0.669 0.669 **
## predorder.m      0.521 0.521 **
## predobjdist.m    0.336 0.336 *
## NOUNcount.m      0.256 0.256 .
## fre             -0.254 0.254 .
## mattr           -0.177 0.177 .
## compoundVERBs   -0.176 0.176 .
## mamr             0.157 0.157 .
## xcomp           -0.114 0.114 .
## NOUNcount.v     -0.111 0.111 .
##
##
## ----- PA5 -----
##               loading abs_l str
## obj             0.897 0.897 ***
## xcomp           0.355 0.355 *
## cli            -0.217 0.217 .
## fre            -0.213 0.213 .
## compoundVERBs  -0.196 0.196 .
## VERBfrac.v     -0.124 0.124 .
## predorder.m     0.100 0.100 .
##
##
## ----- PA4 -----
##                 loading abs_l str
## doubleADPdist.v   0.762 0.762 ***
## doubleADPdist.m   0.681 0.681 **
## subj              0.112 0.112 .
## NOUNcount.v       0.109 0.109 .
```

hypotheses:

- **PA1:** register – narrativity, richness of expression; shorter clauses (-technical / +narrative)
  - narrativity? (1st and 2nd persons etc.)
- **PA2:** text length (-short / +long)
  - hapaxes load negatively, because I normed them over word count
- **PA6:** sentence complexity (more clauses) (-simple / +complex)
  - slightly longer nominal constructions / more objects, more years of education necessary, predicates slightly further in the clause, slightly more verbs
  - `fkgl` in strong correlation with `sentlen.m`
- **PA3:** word length (-short / +long)
  - cli highly correlates with atl, meaning the factor likely expresses mostly token lengths
  - slightly more passives, slightly more objects, slightly less verbal overall / slightly longer nom. constructions, slightly morphologically richer, many years of education necessary
  - more enumerations? but one would expect higher `activity` differences to occur if that was the case
- **PA4:** lexical richness (-poor / +rich)
- **PA5:** passivity (-active / +passive)
  - compound verbs, because that's what passives are in Czech
  - smaller activity, because passive participles count as `ADJ` in UD.

- **PA7:** compound verbs (-less / +more)

strong correlations:

- **PA1–PA6:** (-0.38) narrativity leads to simple clauses
- **PA2–PA6:** (+0.30) longer texts include more complex sentences
- **PA1–PA5:** (-0.49, topconf = +0.09) narrative texts more active

  NOTE: variables with low communalities are excluded from the analysis, yet still likely play a role in legal writing readability. this includes both those selected for the analysis and the excluded ones.

  NOTE: some high-correlating variables were excluded from the FA.

Healthiness diagnostics

```r
fa_res$loadings[] %>%
  as_tibble() %>%
  mutate(feat = colnames(data_scaled)) %>%
  select(feat, everything()) %>%
  pivot_longer(!feat) %>%
  mutate(value = abs(value)) %>%
  group_by(feat) %>%
  summarize(maxload = max(value)) %>%
  arrange(maxload)
```

```
## # A tibble: 21 x 2
##    feat          maxload
##    <chr>           <dbl>
##  1 NEGfrac.m       0.403
##  2 predobjdist.v   0.404
##  3 NOUNcount.v     0.419
##  4 mattr           0.421
##  5 xcomp           0.458
##  6 predobjdist.m   0.489
##  7 predorder.v     0.502
##  8 predorder.m     0.521
##  9 VERBfrac.v      0.590
## 10 cli             0.604
## # i 11 more rows
```

```r
fa_res$communality %>% sort()
```

```
##           NEGfrac.m               mattr          VERBfrac.v       predobjdist.v
##           0.2026778           0.2835459           0.3468350           0.3649748
##         NOUNcount.v               xcomp       predobjdist.m      predsubjdist.m
##           0.4197046           0.4252127           0.4487588           0.4810188
## compoundVERBsdist.m     doubleADPdist.m         predorder.v     doubleADPdist.v
##           0.5054319           0.5220249           0.5665192           0.5764142
##                subj                 cli      predsubjdist.v         predorder.m
##           0.5782792           0.5918487           0.6458289           0.6758268
##       compoundVERBs         NOUNcount.m                mamr                 obj
##           0.6841398           0.7697128           0.7939861           0.8566020
##                 fre
##           0.9894589
```

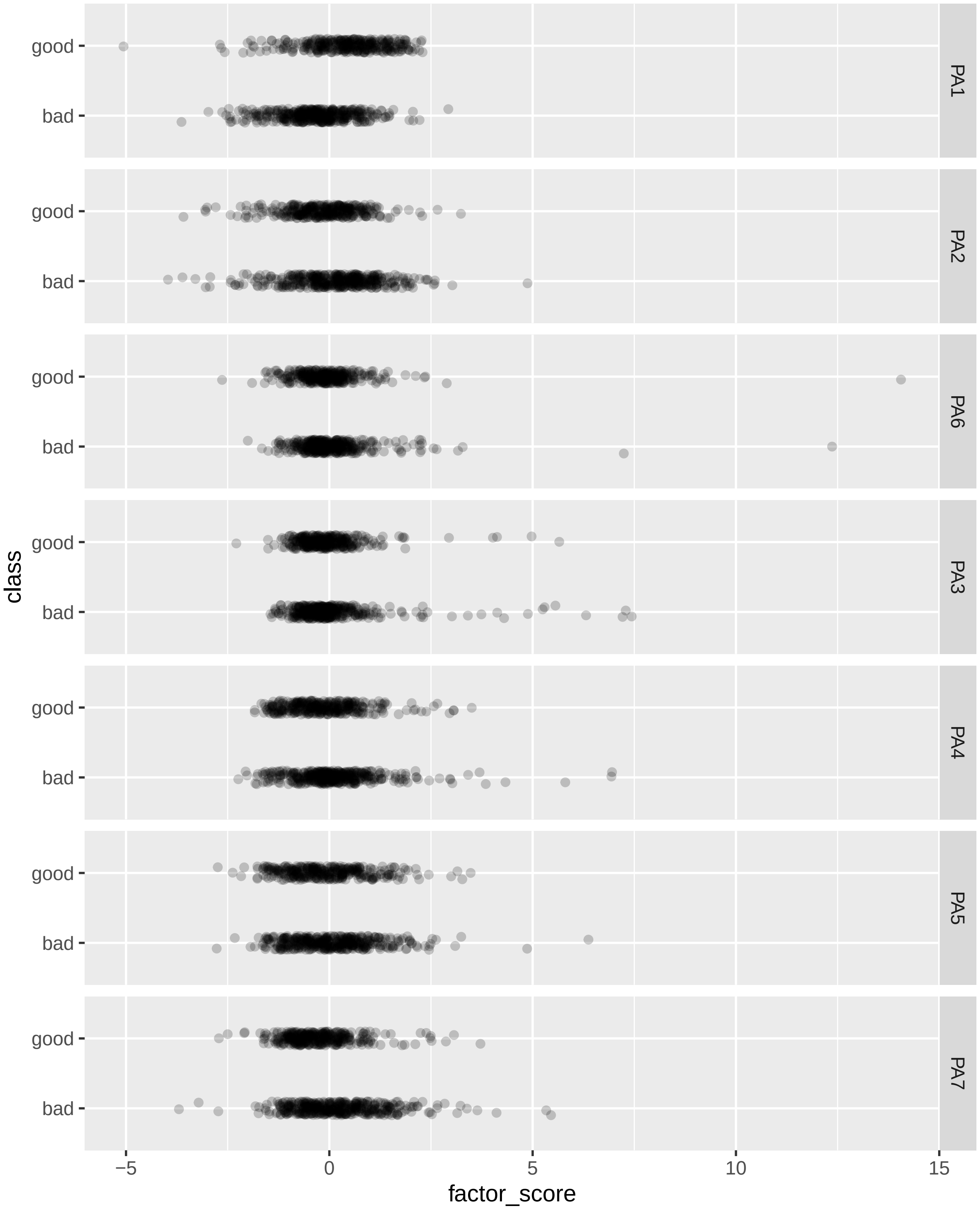
Uniquenesses

```
fa_res$uniquenesses %>% round(3)
```

```
##     doubleADPdist.m     doubleADPdist.v               xcomp       compoundVERBs
##               0.478               0.424               0.575               0.316
## compoundVERBsdist.m         predorder.m         predorder.v                 obj
##               0.495               0.324               0.433               0.143
##       predobjdist.m       predobjdist.v                subj      predsubjdist.m
##               0.551               0.635               0.422               0.519
##      predsubjdist.v          VERBfrac.v           NEGfrac.m         NOUNcount.m
##               0.354               0.653               0.797               0.230
##         NOUNcount.v                 cli                 fre                mamr
##               0.580               0.408               0.011               0.206
##               mattr
##               0.716
```

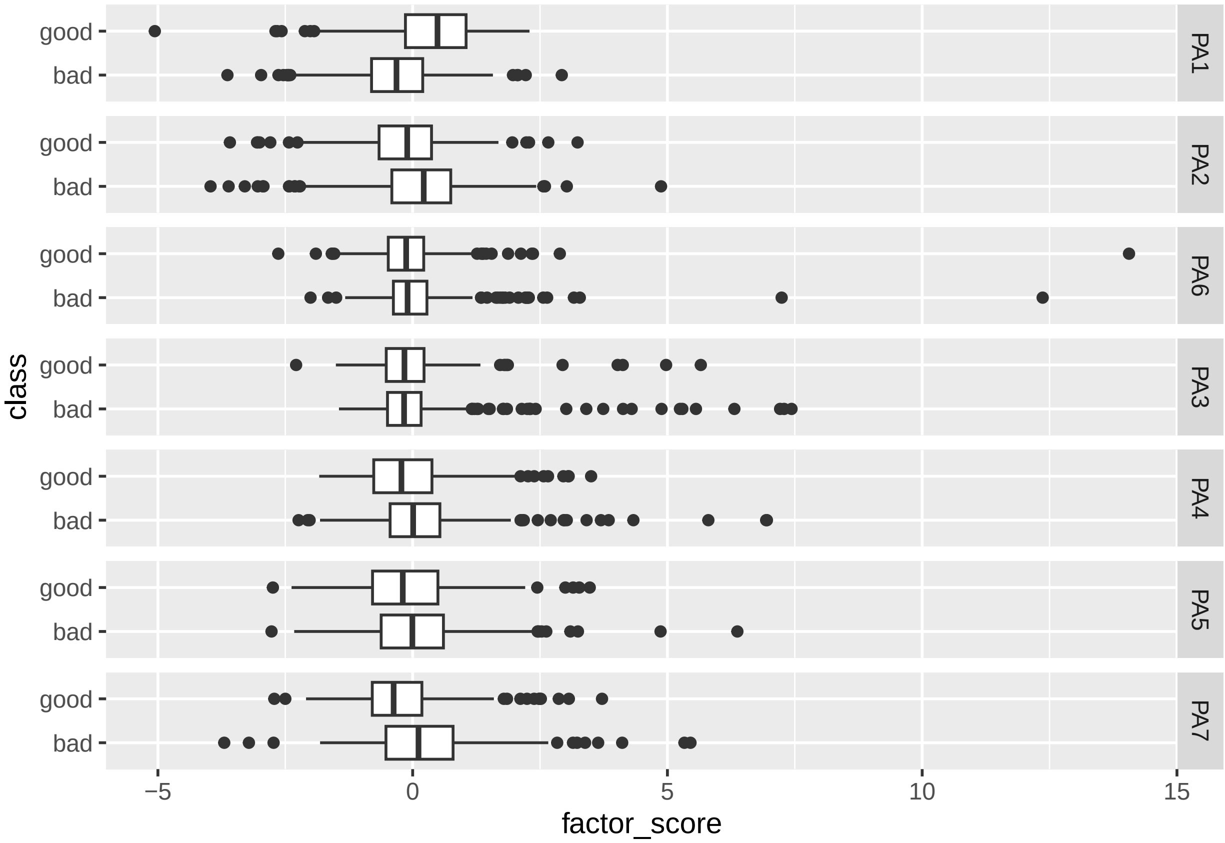
Distributions over factors

```
analyze_distributions <- function(data_factors_long, variable) {
  plot <- data_factors_long %>%
    ggplot(aes(x = factor_score, y = !!sym(variable))) +
    geom_boxplot() +
    facet_grid(factor ~ .)
  print(plot)

  formula <- reformulate(variable, "factor_score")
  factors <- levels(data_factors_long$factor)

  min_p_values <- numeric()
  for (f in factors) {
    data <- data_factors_long %>% filter(factor == f)

    cat(
      "\nTest for the significance of differences in",
      variable, "over", f, ":\n\n"
    )

    dunn <- dunn.test(
      data$factor_score, data[[variable]],
      altp = TRUE, method = "bonferroni"
    )
    min_p_values <- c(min_p_values, min(dunn$altP.adjusted))
  }

  cat(
    "\np < 5e-2\tfound in:",
    factors[min_p_values < 0.05],
    "\np < 1e-2\tfound in:",
    factors[min_p_values < 0.01],
    "\np < 1e-3\tfound in:",
    factors[min_p_values < 0.001],
    "\np < 1e-4\tfound in:",
    factors[min_p_values < 0.0001], "\n"
  )
}

data_factors <- bind_cols(data_pure, fa_res$scores %>% as.data.frame())
colnames(data_factors) <- prettify_feat_name_vector(colnames(data_factors))

data_factors_noout <- bind_cols(data_no_out, fa_res$scores %>% as.data.frame())
colnames(data_factors_noout) <- prettify_feat_name_vector(
  colnames(data_factors_noout)
)

data_factors_long <- data_factors %>%
  pivot_longer(PA1:PA4, names_to = "factor", values_to = "factor_score") %>%
  mutate(across(
    factor,
    ~ factor(.x, levels = c("PA1", "PA2", "PA6", "PA3", "PA4", "PA5", "PA7"))
  ))

data_factors_noout_long <- data_factors_noout %>%
  pivot_longer(PA1:PA4, names_to = "factor", values_to = "factor_score") %>%
  mutate(across(
    factor,
    ~ factor(.x, levels = c("PA1", "PA2", "PA6", "PA3", "PA4", "PA5", "PA7"))
  ))

data_factors_noout_long %>%
  ggplot(aes(x = factor_score, y = class)) +
  facet_grid(factor ~ .) +
  theme(legend.position = "bottom") +
  geom_jitter(width = 0, height = 0.1, alpha = 0.2)
```



class

```
analyze_distributions(data_factors_noout_long, "class")
```



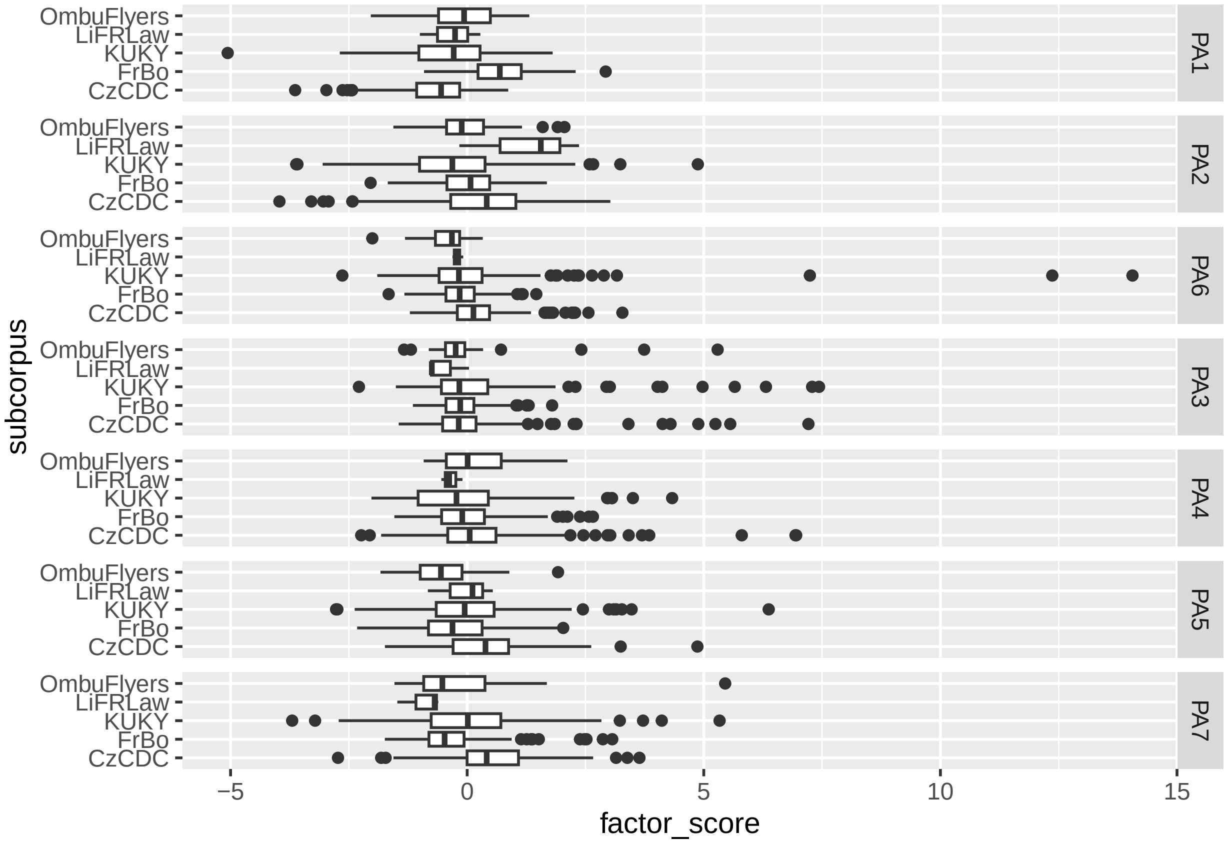
```
##
## Test for the significance of differences in class over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 128.6297, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |        bad
## ---------+-----------
##     good |  -11.34150
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in class over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 22.9771, df = 1, p-value = 0
##
```

```
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |        bad
## ---------+-----------
##     good |   4.793438
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in class over PA6 :
##
##   Kruskal-Wallis rank sum test
##
## data:  x and group
## Kruskal-Wallis chi-squared = 3.5239, df = 1, p-value = 0.06
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |        bad
## ---------+-----------
##     good |   1.877201
##          |     0.0605
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in class over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data:  x and group
## Kruskal-Wallis chi-squared = 0.0145, df = 1, p-value = 0.9
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |        bad
## ---------+-----------
##     good |   0.120387
##          |     0.9042
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in class over PA4 :
##
##   Kruskal-Wallis rank sum test
##
```

```
## data: x and group
## Kruskal-Wallis chi-squared = 14.7309, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |        bad
## ---------+-----------
##     good |   3.838084
##          |    0.0001*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in class over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 4.9509, df = 1, p-value = 0.03
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |        bad
## ---------+-----------
##     good |   2.225061
##          |    0.0261*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in class over PA7 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 48.6256, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |        bad
## ---------+-----------
##     good |   6.973207
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## p < 5e-2 found in: PA1 PA2 PA4 PA5 PA7
## p < 1e-2 found in: PA1 PA2 PA4 PA7
## p < 1e-3 found in: PA1 PA2 PA4 PA7
## p < 1e-4 found in: PA1 PA2 PA7
```

subcorpus

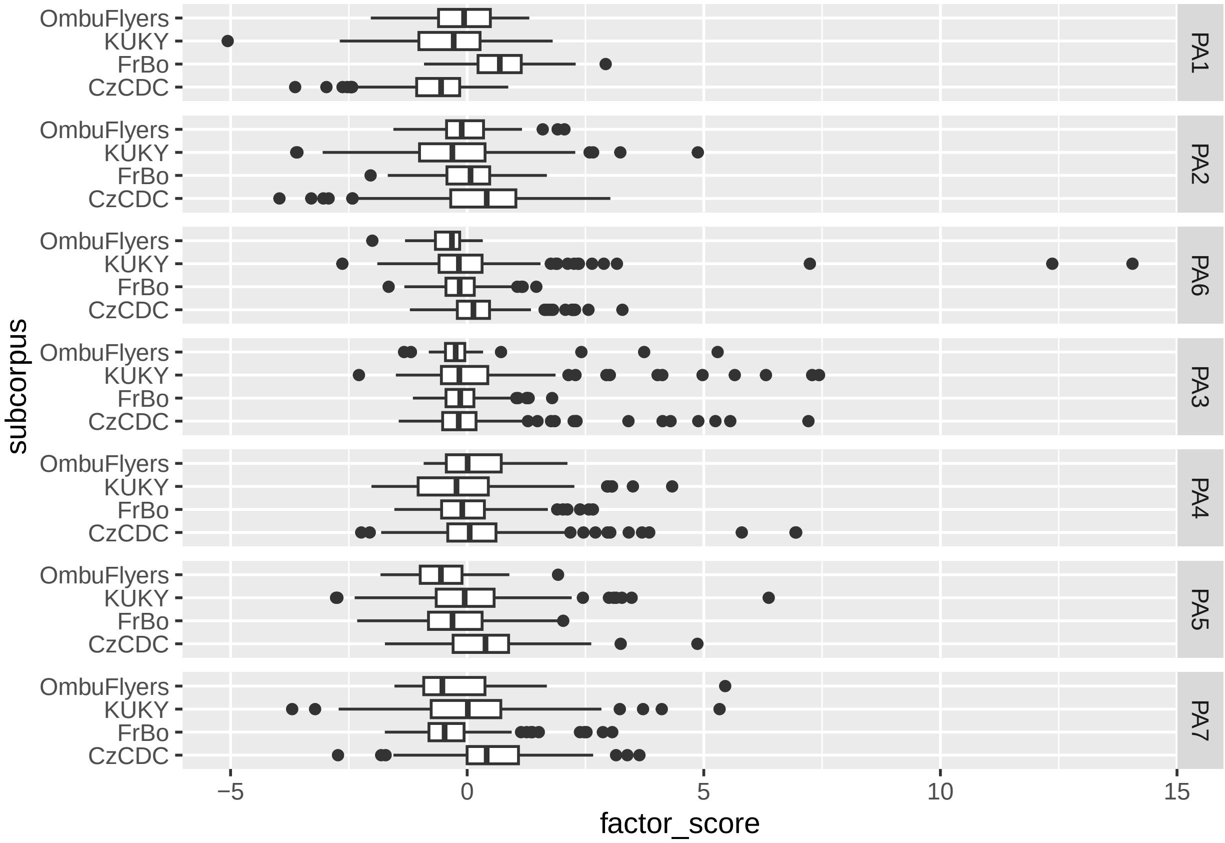
```
analyze_distributions(data_factors_noout_long, "subcorpus")
```



```
##
## Test for the significance of differences in subcorpus over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 301.406, df = 4, p-value = 0
##
##
##                              Comparison of x by group
##                                    (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY     LiFRLaw
## ---------+--------------------------------------------
##     FrBo |  -16.18100
##          |    0.0000*
##          |
##     KUKY |  -3.288885   12.12540
##          |    0.0101*    0.0000*
##          |
##  LiFRLaw |  -0.489587   1.996490   0.073293
##          |     1.0000     0.4588     1.0000
##          |
## OmbuFlye |  -3.390292   4.919012  -1.519189  -0.520883
##          |    0.0070*    0.0000*     1.0000     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in subcorpus over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 44.2049, df = 4, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY    LiFRLaw
## ---------+--------------------------------------------
##     FrBo |   3.477849
##          |    0.0051*
##          |
##     KUKY |   6.368427   3.517208
##          |    0.0000*    0.0044*
##          |
##  LiFRLaw |  -1.146380  -1.683393  -2.234724
##          |     1.0000     0.9230     0.2544
##          |
## OmbuFlye |   2.559757   0.817779  -1.030596   1.863036
##          |     0.1047     1.0000     1.0000     0.6246
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in subcorpus over PA6 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 52.5204, df = 4, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY    LiFRLaw
## ---------+--------------------------------------------
##     FrBo |   5.871330
##          |    0.0000*
##          |
##     KUKY |   4.757207  -0.545201
##          |    0.0000*     1.0000
##          |
##  LiFRLaw |   1.149127   0.249118   0.334614
##          |     1.0000     1.0000     1.0000
##          |
## OmbuFlye |   5.547431   2.636055   2.835711   0.514898
##          |    0.0000*     0.0839    0.0457*     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in subcorpus over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 3.2314, df = 4, p-value = 0.52
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY    LiFRLaw
## ---------+--------------------------------------------
##     FrBo |  -0.453511
##          |     1.0000
##          |
##     KUKY |  -0.768091  -0.391205
##          |     1.0000     1.0000
##          |
##  LiFRLaw |   1.195977   1.268236   1.326439
##          |     1.0000     1.0000     1.0000
##          |
## OmbuFlye |   0.761529   1.014988   1.185807  -0.935855
##          |     1.0000     1.0000     1.0000     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in subcorpus over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 18.7694, df = 4, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY    LiFRLaw
## ---------+--------------------------------------------
##     FrBo |   2.257323
##          |      0.2399
##          |
##     KUKY |    3.869033   1.996858
##          |     0.0011*     0.4584
##          |
##  LiFRLaw |    1.125516   0.780952   0.462928
##          |      1.0000     1.0000     1.0000
##          |
## OmbuFlye |   -0.569874  -1.754104  -2.733757  -1.258499
##          |      1.0000     0.7941     0.0626     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in subcorpus over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 58.0389, df = 4, p-value = 0
##
##
##                             Comparison of x by group
##                                   (Bonferroni)
## Col Mean-|
## Row Mean |       CzCDC       FrBo       KUKY    LiFRLaw
## ---------+---------------------------------------------
##     FrBo |    6.810837
##          |     0.0000*
##          |
##     KUKY |    3.281456  -3.051904
##          |     0.0103*    0.0227*
##          |
##  LiFRLaw |    0.643168  -0.402328   0.081441
##          |      1.0000     1.0000     1.0000
##          |
## OmbuFlye |    5.194442   1.787403   3.312748   0.901761
##          |     0.0000*     0.7387    0.0092*     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in subcorpus over PA7 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 148.2076, df = 4, p-value = 0
##
##
##                             Comparison of x by group
##                                   (Bonferroni)
## Col Mean-|
## Row Mean |       CzCDC       FrBo       KUKY    LiFRLaw
## ---------+--------------------------------------------
##     FrBo |   11.89715
##          |    0.0000*
##          |
##     KUKY |   5.780043  -5.279138
##          |    0.0000*    0.0000*
##          |
##  LiFRLaw |   2.979883   1.157572   1.989041
##          |    0.0288*     1.0000     0.4670
##          |
## OmbuFlye |   5.015208  -1.033665   1.734554  -1.416257
##          |    0.0000*     1.0000     0.8282     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## p < 5e-2 found in: PA1 PA2 PA6 PA4 PA5 PA7
## p < 1e-2 found in: PA1 PA2 PA6 PA4 PA5 PA7
## p < 1e-3 found in: PA1 PA2 PA6 PA5 PA7
## p < 1e-4 found in: PA1 PA2 PA6 PA5 PA7
```

subcorpus wo/ LiFRLaw

```
analyze_distributions(
  data_factors_noout_long %>% filter(subcorpus != "LiFRLaw"), "subcorpus"
)
```



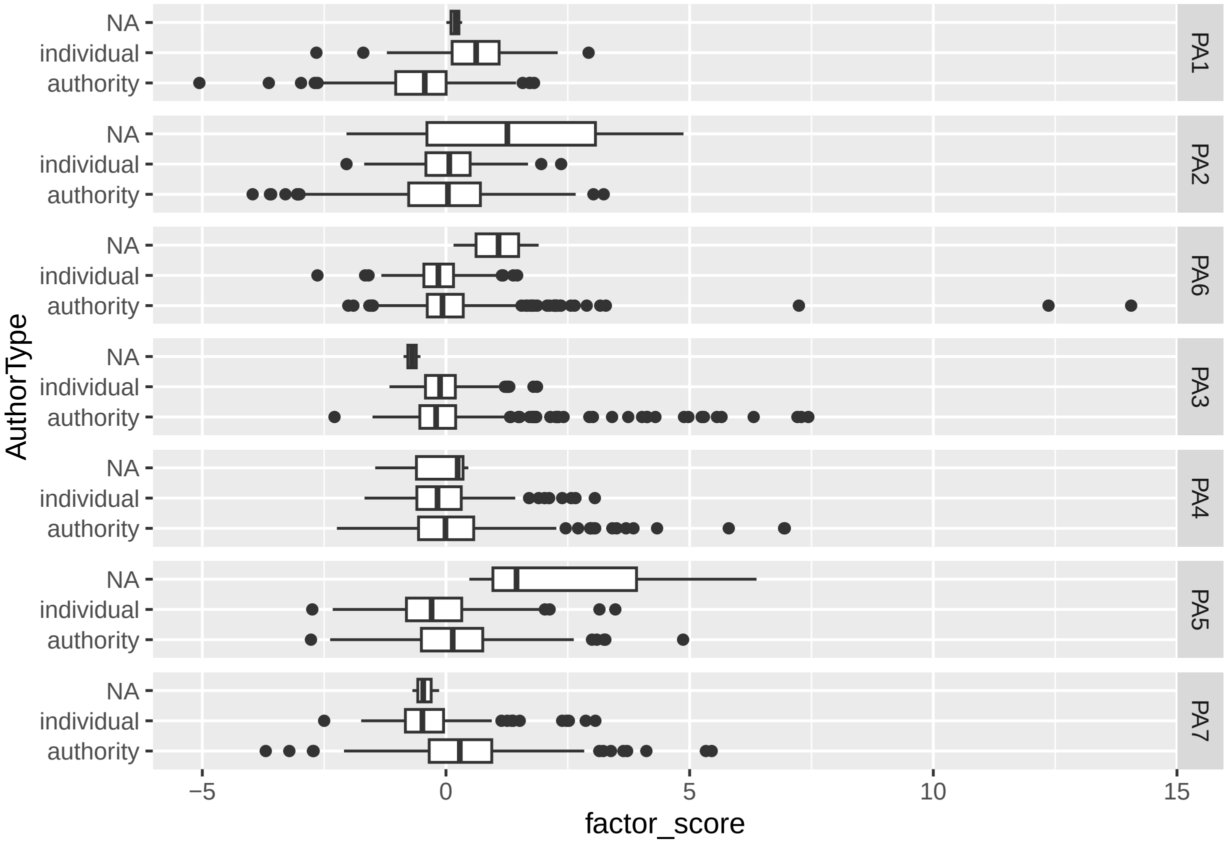
```
##
## Test for the significance of differences in subcorpus over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 300.3341, df = 3, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY
## ---------+---------------------------------
##     FrBo |  -16.16644
##          |    0.0000*
##          |
##     KUKY |  -3.286217   12.11418
##          |    0.0061*    0.0000*
##          |
## OmbuFlye |  -3.386917   4.914920  -1.517337
##          |    0.0042*    0.0000*     0.7751
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in subcorpus over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 41.4788, df = 3, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY
## ---------+---------------------------------
##     FrBo |   3.481111
##          |    0.0030*
##          |
##     KUKY |   6.378773   3.525237
##          |    0.0000*    0.0025*
##          |
## OmbuFlye |   2.564073   0.820507  -1.032115
##          |     0.0621     1.0000     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in subcorpus over PA6 :
##
##   Kruskal-Wallis rank sum test
```

```
##
## data: x and group
## Kruskal-Wallis chi-squared = 52.1067, df = 3, p-value = 0
##
##
##                           Comparison of x by group
##                                 (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY
## ---------+---------------------------------
##     FrBo |   5.863896
##          |    0.0000*
##          |
##     KUKY |   4.746639  -0.549426
##          |    0.0000*     1.0000
##          |
## OmbuFlye |   5.537045   2.629274   2.831333
##          |    0.0000*     0.0513    0.0278*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in subcorpus over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 1.6668, df = 3, p-value = 0.64
##
##
##                           Comparison of x by group
##                                 (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY
## ---------+---------------------------------
##     FrBo |  -0.449994
##          |     1.0000
##          |
##     KUKY |  -0.770950  -0.397707
##          |     1.0000     1.0000
##          |
## OmbuFlye |   0.762235   1.013887   1.188109
##          |     1.0000     1.0000     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in subcorpus over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 18.0035, df = 3, p-value = 0
##
```

```
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY
## ---------+---------------------------------
##     FrBo |   2.253775
##          |     0.1453
##          |
##     KUKY |   3.858712   1.989133
##          |    0.0007*     0.2801
##          |
## OmbuFlye |  -0.571384  -1.753811  -2.729470
##          |     1.0000     0.4768    0.0381*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in subcorpus over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 57.9248, df = 3, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY
## ---------+---------------------------------
##     FrBo |   6.803584
##          |    0.0000*
##          |
##     KUKY |   3.276050  -3.050721
##          |    0.0063*    0.0137*
##          |
## OmbuFlye |   5.189707   1.786316   3.311082
##          |    0.0000*     0.4443    0.0056*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in subcorpus over PA7 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 144.995, df = 3, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY
## ---------+---------------------------------
##     FrBo |   11.91522
##          |    0.0000*
##          |
##     KUKY |   5.780161  -5.296526
##          |    0.0000*    0.0000*
##          |
## OmbuFlye |   5.019190  -1.038959   1.738437
##          |    0.0000*     1.0000     0.4928
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## p < 5e-2 found in: PA1 PA2 PA6 PA4 PA5 PA7
## p < 1e-2 found in: PA1 PA2 PA6 PA4 PA5 PA7
## p < 1e-3 found in: PA1 PA2 PA6 PA4 PA5 PA7
## p < 1e-4 found in: PA1 PA2 PA6 PA5 PA7
```

AuthorType

```
analyze_distributions(data_factors_noout_long, "AuthorType")
```



```
##
## Test for the significance of differences in AuthorType over PA1 :
##
##   Kruskal-Wallis rank sum test
```

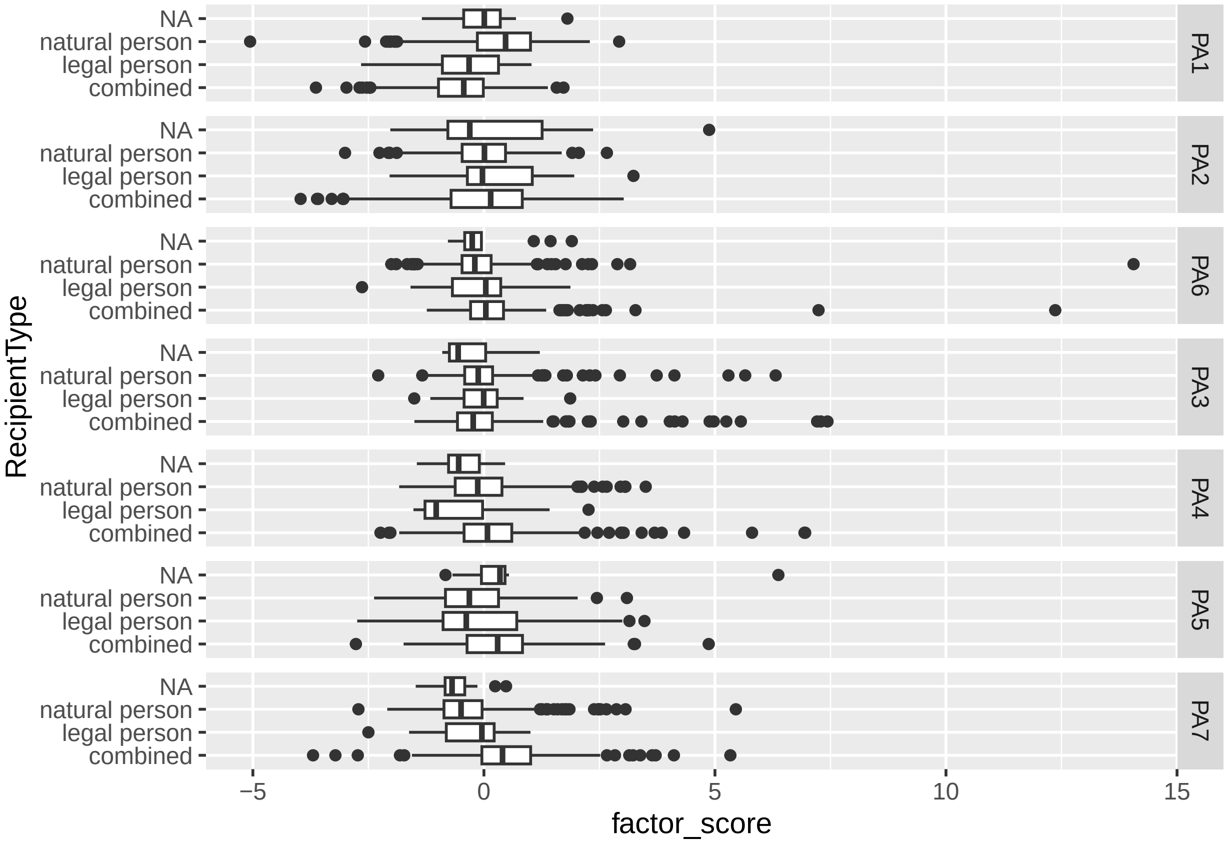
```
##
## data: x and group
## Kruskal-Wallis chi-squared = 273.5423, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## ---------+-----------
## individu |  -16.53911
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in AuthorType over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.2839, df = 1, p-value = 0.59
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## ---------+-----------
## individu |  -0.532864
##          |     0.5941
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in AuthorType over PA6 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 10.8429, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## ---------+-----------
## individu |   3.292854
##          |    0.0010*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
```

```
## Test for the significance of differences in AuthorType over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 2.01, df = 1, p-value = 0.16
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## ---------+-----------
## individu |  -1.417754
##          |     0.1563
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in AuthorType over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 5.31, df = 1, p-value = 0.02
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## ---------+-----------
## individu |   2.304337
##          |    0.0212*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in AuthorType over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 25.2579, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## ---------+-----------
## individu |   5.025721
##          |    0.0000*
##
```

```
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in AuthorType over PA7 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 123.5727, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## ---------+-----------
## individu |   11.11632
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## p < 5e-2 found in: PA1 PA6 PA4 PA5 PA7
## p < 1e-2 found in: PA1 PA6 PA5 PA7
## p < 1e-3 found in: PA1 PA6 PA5 PA7
## p < 1e-4 found in: PA1 PA5 PA7
```

RecipientType

```
analyze_distributions(data_factors_noout_long, "RecipientType")
```



```
##
## Test for the significance of differences in RecipientType over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 185.075, df = 2, p-value = 0
##
##
##                              Comparison of x by group
##                                    (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## ---------+----------------------
## legal pe |  -1.139892
##          |     0.7630
##          |
## natural  |  -13.51017  -3.610577
##          |    0.0000*    0.0009*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in RecipientType over PA2 :
##
##   Kruskal-Wallis rank sum test
##
```

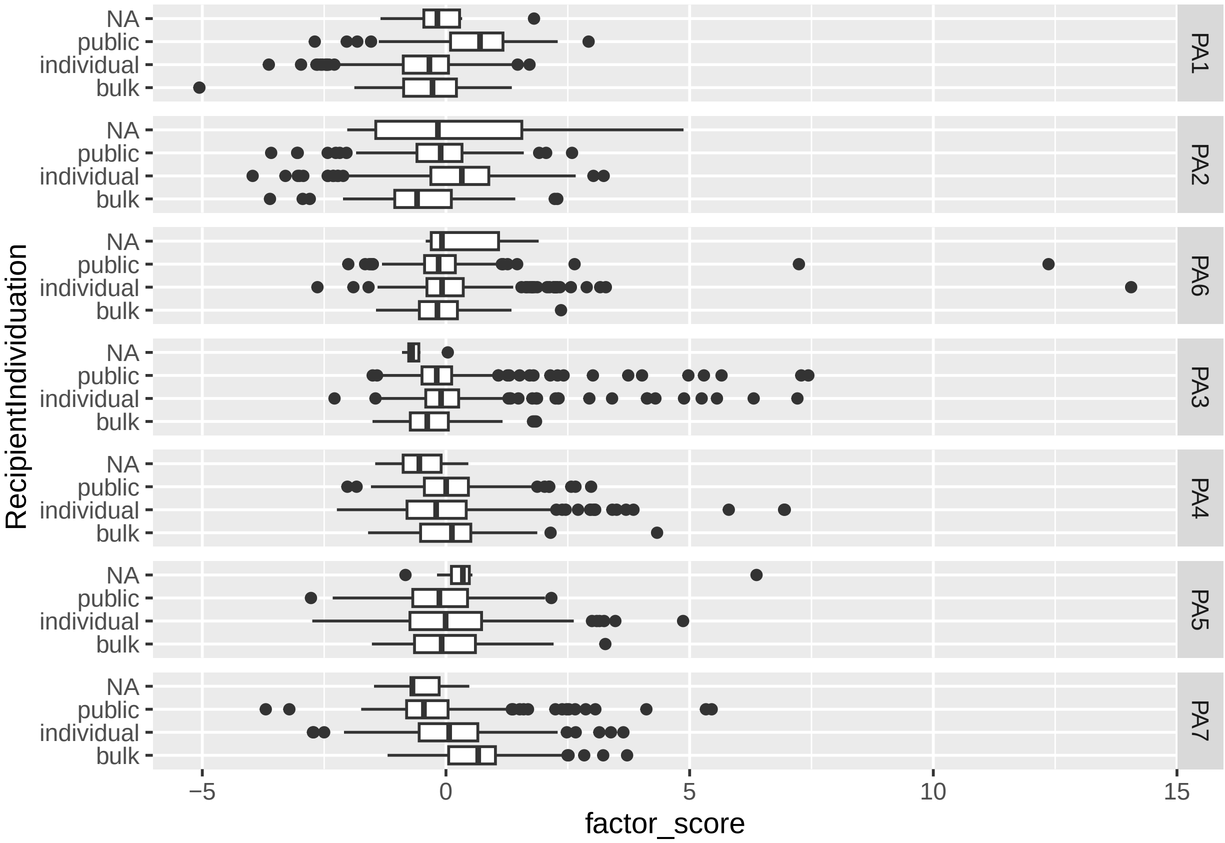
```
## data: x and group
## Kruskal-Wallis chi-squared = 2.5567, df = 2, p-value = 0.28
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## ---------+----------------------
## legal pe |  -0.480884
##          |     1.0000
##          |
## natural  |   1.397820   0.977808
##          |     0.4865     0.9845
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in RecipientType over PA6 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 32.9674, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## ---------+----------------------
## legal pe |   1.369390
##          |     0.5126
##          |
## natural  |   5.737297   0.639954
##          |    0.0000*     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in RecipientType over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 5.0437, df = 2, p-value = 0.08
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## ---------+----------------------
## legal pe |  -1.286861
##          |     0.5944
##          |
## natural  |  -2.060919   0.572271
##          |     0.1179     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in RecipientType over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 17.7411, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## ---------+----------------------
## legal pe |   3.459097
##          |    0.0016*
##          |
## natural  |   3.047082  -2.416659
##          |    0.0069*    0.0470*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in RecipientType over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 47.5454, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## ---------+----------------------
## legal pe |   2.065791
##          |     0.1165
##          |
## natural  |   6.858974   0.332512
##          |    0.0000*     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in RecipientType over PA7 :
##
```

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 192.2065, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## ---------+----------------------
## legal pe |   3.980224
##          |    0.0002*
##          |
## natural  |   13.80818   0.849552
##          |    0.0000*     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## p < 5e-2 found in: PA1 PA6 PA4 PA5 PA7
## p < 1e-2 found in: PA1 PA6 PA4 PA5 PA7
## p < 1e-3 found in: PA1 PA6 PA5 PA7
## p < 1e-4 found in: PA1 PA6 PA5 PA7
```

court decisions often with `RecipientType = combined`.

RecipientIndividuation

```
analyze_distributions(data_factors_noout_long, "RecipientIndividuation")
```



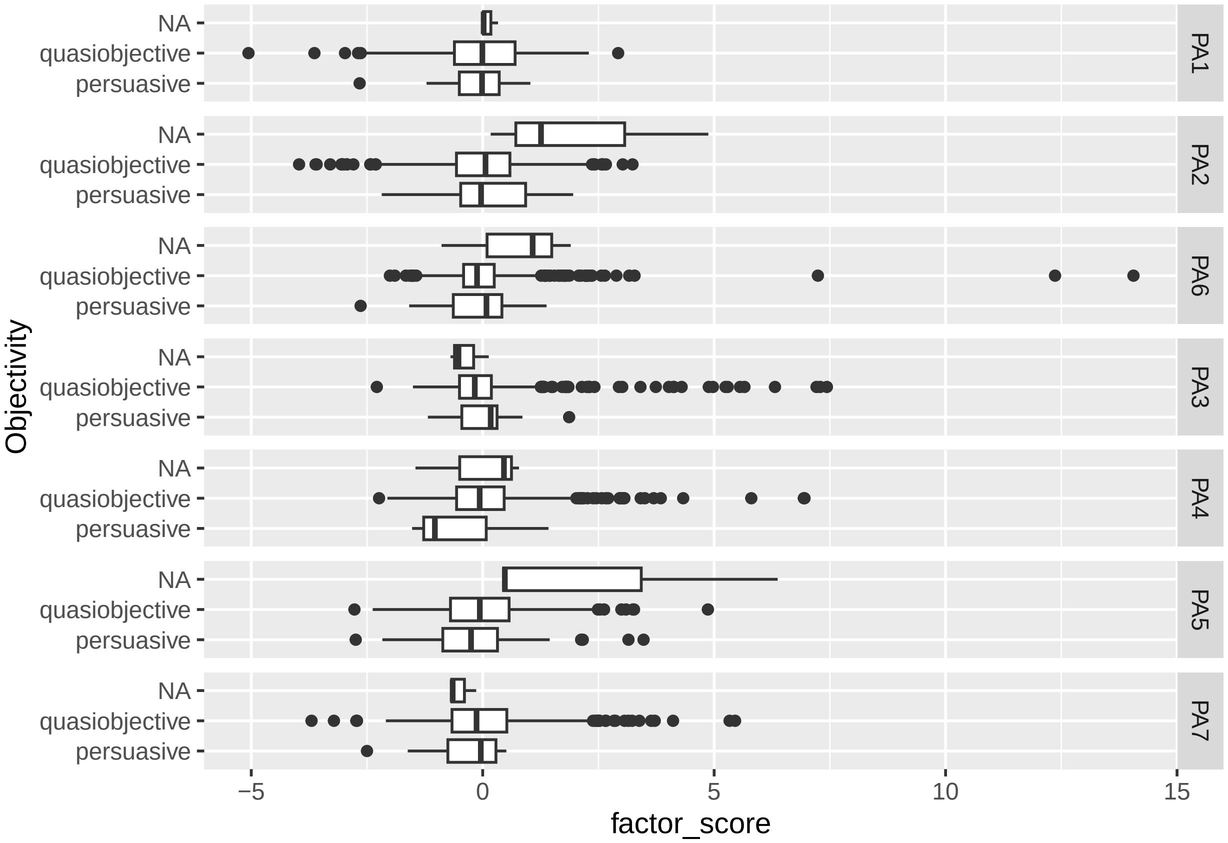
```
##
## Test for the significance of differences in RecipientIndividuation over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 225.8919, df = 2, p-value = 0
##
##
##                              Comparison of x by group
##                                    (Bonferroni)
## Col Mean-|
## Row Mean |       bulk   individu
## ---------+----------------------
## individu |   0.733244
##          |     1.0000
##          |
##   public |  -7.769489  -14.63294
##          |    0.0000*    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in RecipientIndividuation over PA2 :
##
##   Kruskal-Wallis rank sum test
##
```

```
## data: x and group
## Kruskal-Wallis chi-squared = 56.9616, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |       bulk   individu
## ---------+----------------------
## individu |  -6.061693
##          |    0.0000*
##          |
##   public |  -2.520497   5.992273
##          |    0.0352*    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in RecipientIndividuation over PA6 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 8.8962, df = 2, p-value = 0.01
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |       bulk   individu
## ---------+----------------------
## individu |  -1.880023
##          |     0.1803
##          |
##   public |  -0.267727   2.743856
##          |     1.0000    0.0182*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in RecipientIndividuation over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 14.9619, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |       bulk   individu
## ---------+----------------------
## individu |  -3.732691
##          |    0.0006*
##          |
##   public |  -2.531216   2.003384
##          |    0.0341*     0.1354
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in RecipientIndividuation over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 13.1559, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |       bulk   individu
## ---------+----------------------
## individu |   2.355550
##          |     0.0555
##          |
##   public |   0.415310  -3.300313
##          |     1.0000    0.0029*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in RecipientIndividuation over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 3.7625, df = 2, p-value = 0.15
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |       bulk   individu
## ---------+----------------------
## individu |   0.045335
##          |     1.0000
##          |
##   public |   1.119860   1.851671
##          |     0.7883     0.1922
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in RecipientIndividuation over PA7 :
##
```

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 80.5375, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |       bulk   individu
## ---------+----------------------
## individu |   3.974094
##          |    0.0002*
##          |
##   public |   7.788218   6.640251
##          |    0.0000*    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## p < 5e-2 found in: PA1 PA2 PA6 PA3 PA4 PA7
## p < 1e-2 found in: PA1 PA2 PA3 PA4 PA7
## p < 1e-3 found in: PA1 PA2 PA3 PA7
## p < 1e-4 found in: PA1 PA2 PA7
```

Objectivity

```
analyze_distributions(data_factors_noout_long, "Objectivity")
```



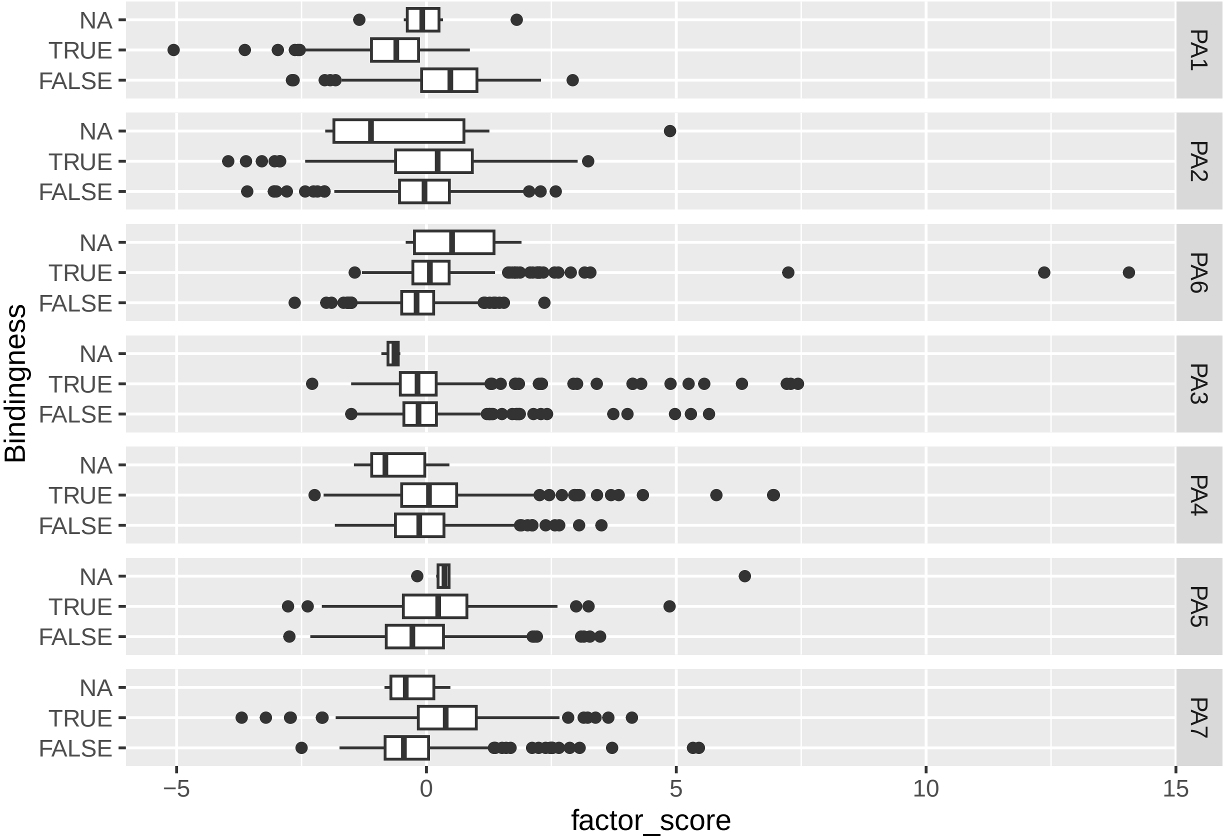
```
##
## Test for the significance of differences in Objectivity over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.689, df = 1, p-value = 0.41
##
##
##                              Comparison of x by group
##                                    (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## ---------+-----------
## quasiobj |  -0.830081
##          |     0.4065
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in Objectivity over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.0936, df = 1, p-value = 0.76
##
```

```
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## ---------+-----------
## quasiobj |   0.305980
##          |     0.7596
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in Objectivity over PA6 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.4976, df = 1, p-value = 0.48
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## ---------+-----------
## quasiobj |   0.705441
##          |     0.4805
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in Objectivity over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.9569, df = 1, p-value = 0.33
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## ---------+-----------
## quasiobj |   0.978218
##          |     0.3280
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in Objectivity over PA4 :
##
##   Kruskal-Wallis rank sum test
##
```

```
## data: x and group
## Kruskal-Wallis chi-squared = 9.4784, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## ---------+-----------
## quasiobj |  -3.078707
##          |    0.0021*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in Objectivity over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.1624, df = 1, p-value = 0.69
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## ---------+-----------
## quasiobj |  -0.403036
##          |     0.6869
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in Objectivity over PA7 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.937, df = 1, p-value = 0.33
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## ---------+-----------
## quasiobj |  -0.968002
##          |     0.3330
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## p < 5e-2 found in: PA4
## p < 1e-2 found in: PA4
## p < 1e-3 found in:
## p < 1e-4 found in:
```

Bindingness

```
analyze_distributions(data_factors_noout_long, "Bindingness")
```



```
##
## Test for the significance of differences in Bindingness over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 265.5128, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## ---------+-----------
##     TRUE |   16.29456
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
```

```
##
## Test for the significance of differences in Bindingness over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 9.5665, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## ---------+-----------
##     TRUE |  -3.092982
##          |    0.0020*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in Bindingness over PA6 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 44.6818, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## ---------+-----------
##     TRUE |  -6.684444
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in Bindingness over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.2572, df = 1, p-value = 0.61
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## ---------+-----------
##     TRUE |   0.507102
##          |     0.6121
```

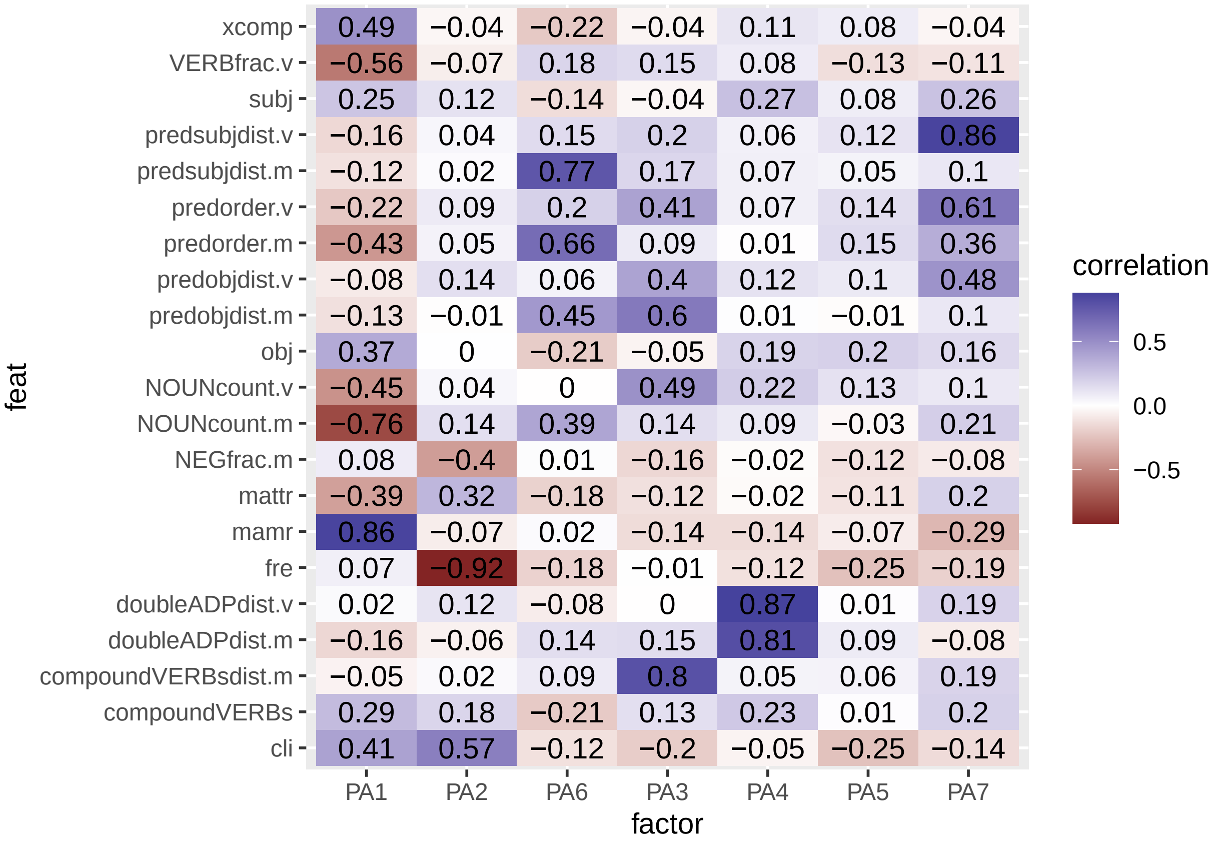
```
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in Bindingness over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 8.321, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## ---------+-----------
##     TRUE |  -2.884622
##          |    0.0039*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in Bindingness over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 31.4668, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## ---------+-----------
##     TRUE |  -5.609524
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## Test for the significance of differences in Bindingness over PA7 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 134.0698, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## ---------+-----------
##     TRUE |  -11.57885
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
##
## p < 5e-2 found in: PA1 PA2 PA6 PA4 PA5 PA7
## p < 1e-2 found in: PA1 PA2 PA6 PA4 PA5 PA7
## p < 1e-3 found in: PA1 PA6 PA5 PA7
## p < 1e-4 found in: PA1 PA6 PA5 PA7
```

Feature-factor correlations

```r
data_factors_clean_longer <- data_factors_noout_long %>%
  pivot_longer(
    abstractNOUNs:verbdist,
    names_to = "feat", values_to = "feat_value"
  )

data_factors_correlations <- data_factors_clean_longer %>%
  group_by(feat, factor) %>%
  summarize(correlation = cor(feat_value, factor_score))
```

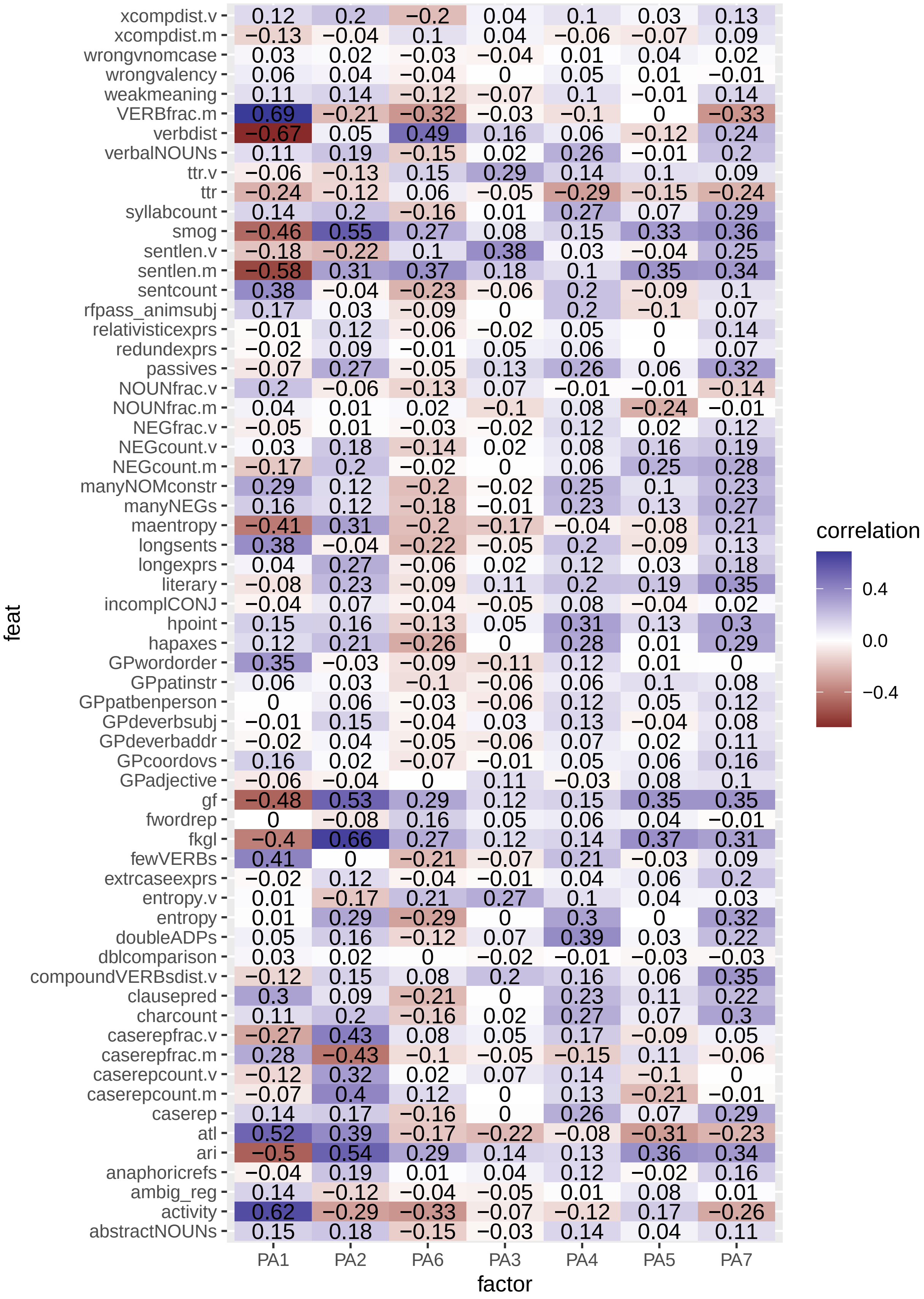
```
## `summarise()` has grouped output by 'feat'. You can override using the
## `.groups` argument.
```

```r
data_factors_correlations %>%
  filter(feat %in% final_collist) %>%
  ggplot(aes(
    x = factor,
    y = feat,
    fill = correlation,
    label = round(correlation, 2)
  )) +
  geom_tile() +
  geom_text() +
  scale_fill_gradient2()
```



feat	PA1	PA2	PA6	PA3	PA4	PA5	PA7
xcomp	0.49	−0.04	−0.22	−0.04	0.11	0.08	−0.04
VERBfrac.v	−0.56	−0.07	0.18	0.15	0.08	−0.13	−0.11
subj	0.25	0.12	−0.14	−0.04	0.27	0.08	0.26
predsubjdist.v	−0.16	0.04	0.15	0.2	0.06	0.12	0.86
predsubjdist.m	−0.12	0.02	0.77	0.17	0.07	0.05	0.1
predorder.v	−0.22	0.09	0.2	0.41	0.07	0.14	0.61
predorder.m	−0.43	0.05	0.66	0.09	0.01	0.15	0.36
predobjdist.v	−0.08	0.14	0.06	0.4	0.12	0.1	0.48
predobjdist.m	−0.13	−0.01	0.45	0.6	0.01	−0.01	0.1
obj	0.37	0	−0.21	−0.05	0.19	0.2	0.16
NOUNcount.v	−0.45	0.04	0	0.49	0.22	0.13	0.1
NOUNcount.m	−0.76	0.14	0.39	0.14	0.09	−0.03	0.21
NEGfrac.m	0.08	−0.4	0.01	−0.16	−0.02	−0.12	−0.08
mattr	−0.39	0.32	−0.18	−0.12	−0.02	−0.11	0.2
mamr	0.86	−0.07	0.02	−0.14	−0.14	−0.07	−0.29
fre	0.07	−0.92	−0.18	−0.01	−0.12	−0.25	−0.19
doubleADPdist.v	0.02	0.12	−0.08	0	0.87	0.01	0.19
doubleADPdist.m	−0.16	−0.06	0.14	0.15	0.81	0.09	−0.08
compoundVERBsdist.m	−0.05	0.02	0.09	0.8	0.05	0.06	0.19
compoundVERBs	0.29	0.18	−0.21	0.13	0.23	0.01	0.2
cli	0.41	0.57	−0.12	−0.2	−0.05	−0.25	−0.14

```r
data_factors_correlations %>%
  filter(!(feat %in% final_collist)) %>%
  ggplot(aes(
    x = factor,
    y = feat,
    fill = correlation,
    label = round(correlation, 2)
  )) +
  geom_tile() +
  geom_text() +
  scale_fill_gradient2()
```



feat	PA1	PA2	PA6	PA3	PA4	PA5	PA7
xcompdist.v	0.12	0.2	−0.2	0.04	0.1	0.03	0.13
xcompdist.m	−0.13	−0.04	0.1	0.04	−0.06	−0.07	0.09
wrongvnomcase	0.03	0.02	−0.03	−0.04	0.01	0.04	0.02
wrongvalency	0.06	0.04	−0.04	0	0.05	0.01	−0.01
weakmeaning	0.11	0.14	−0.12	−0.07	0.1	−0.01	0.14
VERBfrac.m	0.69	−0.21	−0.32	−0.03	−0.1	0	−0.33
verbdist	−0.67	0.05	0.49	0.16	0.06	−0.12	0.24
verbalNOUNs	0.11	0.19	−0.15	0.02	0.26	−0.01	0.2
ttr.v	−0.06	−0.13	0.15	0.29	0.14	0.1	0.09
ttr	−0.24	−0.12	0.06	−0.05	−0.29	−0.15	−0.24
syllabcount	0.14	0.2	−0.16	0.01	0.27	0.07	0.29
smog	−0.46	0.55	0.27	0.08	0.15	0.33	0.36
sentlen.v	−0.18	−0.22	0.1	0.38	0.03	−0.04	0.25
sentlen.m	−0.58	0.31	0.37	0.18	0.1	0.35	0.34
sentcount	0.38	−0.04	−0.23	−0.06	0.2	−0.09	0.1
rfpass_animsubj	0.17	0.03	−0.09	0	0.2	−0.1	0.07
relativisticexprs	−0.01	0.12	−0.06	−0.02	0.05	0	0.14
redundexprs	−0.02	0.09	−0.01	0.05	0.06	0	0.07
passives	−0.07	0.27	−0.05	0.13	0.26	0.06	0.32
NOUNfrac.v	0.2	−0.06	−0.13	0.07	−0.01	−0.01	−0.14
NOUNfrac.m	0.04	0.01	0.02	−0.1	0.08	−0.24	−0.01
NEGfrac.v	−0.05	0.01	−0.03	−0.02	0.12	0.02	0.12
NEGcount.v	0.03	0.18	−0.14	0.02	0.08	0.16	0.19
NEGcount.m	−0.17	0.2	−0.02	0	0.06	0.25	0.28
manyNOMconstr	0.29	0.12	−0.2	−0.02	0.25	0.1	0.23
manyNEGs	0.16	0.12	−0.18	−0.01	0.23	0.13	0.27
maentropy	−0.41	0.31	−0.2	−0.17	−0.04	−0.08	0.21
longsents	0.38	−0.04	−0.22	−0.05	0.2	−0.09	0.13
longexprs	0.04	0.27	−0.06	0.02	0.12	0.03	0.18
literary	−0.08	0.23	−0.09	0.11	0.2	0.19	0.35
incomplCONJ	−0.04	0.07	−0.04	−0.05	0.08	−0.04	0.02
hpoint	0.15	0.16	−0.13	0.05	0.31	0.13	0.3
hapaxes	0.12	0.21	−0.26	0	0.28	0.01	0.29
GPwordorder	0.35	−0.03	−0.09	−0.11	0.12	0.01	0
GPpatinstr	0.06	0.03	−0.1	−0.06	0.06	0.1	0.08
GPpatbenperson	0	0.06	−0.03	−0.06	0.12	0.05	0.12
GPdeverbsubj	−0.01	0.15	−0.04	0.03	0.13	−0.04	0.08
GPdeverbaddr	−0.02	0.04	−0.05	−0.06	0.07	0.02	0.11
GPcoordovs	0.16	0.02	−0.07	−0.01	0.05	0.06	0.16
GPadjective	−0.06	−0.04	0	0.11	−0.03	0.08	0.1
gf	−0.48	0.53	0.29	0.12	0.15	0.35	0.35
fwordrep	0	−0.08	0.16	0.05	0.06	0.04	−0.01
fkgl	−0.4	0.66	0.27	0.12	0.14	0.37	0.31
fewVERBs	0.41	0	−0.21	−0.07	0.21	−0.03	0.09
extrcaseexprs	−0.02	0.12	−0.04	−0.01	0.04	0.06	0.2
entropy.v	0.01	−0.17	0.21	0.27	0.1	0.04	0.03
entropy	0.01	0.29	−0.29	0	0.3	0	0.32
doubleADPs	0.05	0.16	−0.12	0.07	0.39	0.03	0.22
dblcomparison	0.03	0.02	0	−0.02	−0.01	−0.03	−0.03
compoundVERBsdist.v	−0.12	0.15	0.08	0.2	0.16	0.06	0.35
clausepred	0.3	0.09	−0.21	0	0.23	0.11	0.22
charcount	0.11	0.2	−0.16	0.02	0.27	0.07	0.3
caserepfrac.v	−0.27	0.43	0.08	0.05	0.17	−0.09	0.05
caserepfrac.m	0.28	−0.43	−0.1	−0.05	−0.15	0.11	−0.06
caserepcount.v	−0.12	0.32	0.02	0.07	0.14	−0.1	0
caserepcount.m	−0.07	0.4	0.12	0	0.13	−0.21	−0.01
caserep	0.14	0.17	−0.16	0	0.26	0.07	0.29
atl	0.52	0.39	−0.17	−0.22	−0.08	−0.31	−0.23
ari	−0.5	0.54	0.29	0.14	0.13	0.36	0.34
anaphoricrefs	−0.04	0.19	0.01	0.04	0.12	−0.02	0.16
ambig_reg	0.14	−0.12	−0.04	−0.05	0.01	0.08	0.01
activity	0.62	−0.29	−0.33	−0.07	−0.12	0.17	−0.26
abstractNOUNs	0.15	0.18	−0.15	−0.03	0.14	0.04	0.11