EFA

```r
set.seed(42)

library(rcompanion) # effect size calculation
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(corrplot)
```

```
## corrplot 0.95 loaded
```

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

```
## Loading required package: boot
```

```
## 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
```

```
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
```

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

```
##
## Attaching package: 'psych'
## The following object is masked from 'package:boot':
##
##     logit
## The following object is masked from 'package:rcompanion':
##
##     phi
```

```
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
```

```
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
```

```
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.
```

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

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

Load and tidy data

```
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.
```

```
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.
```

```
.firstnonmetacolumn <- 17
```

```r
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.v = 0,
    RuleAmbiguousRegards = 0,
    RuleReflexivePassWithAnimSubj = 0,
    RuleTooManyNegations = 0,
    RuleTooManyNegations.max_negation_frac.v = 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.v = 0,
  RulePredObjDistance = 0,
  RulePredObjDistance.max_distance.v = 0,
  RuleInfVerbDistance = 0,
  RuleInfVerbDistance.max_distance.v = 0,
  RuleMultiPartVerbs = 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
)) %>%
# replace NAs with medians
mutate(across(c(
  RuleDoubleAdpos.max_allowable_distance,
  RuleTooManyNegations.max_negation_frac,
  RuleTooManyNegations.max_allowable_negations,
  RulePredSubjDistance.max_distance,
  RulePredObjDistance.max_distance,
  RuleInfVerbDistance.max_distance,
  RuleMultiPartVerbs.max_distance
), ~ coalesce(., median(., na.rm = TRUE)))) %>%
# merge GPs
mutate(
  GPs = RuleGPcoordovs +
    RuleGPdeverbaddr +
    RuleGPpatinstr +
    RuleGPdeverbsubj +
    RuleGPadjective +
    RuleGPpatbenperson +
    RuleGPwordorder
) %>%
select(!c(
  RuleGPcoordovs,
  RuleGPdeverbaddr,
  RuleGPpatinstr,
  RuleGPdeverbsubj,
  RuleGPadjective,
  RuleGPpatbenperson,
  RuleGPwordorder
))
```

```
data_clean <- data_no_nas %>%
  # norm data expected to correlate with text length
  mutate(across(c(
    GPs,
    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,
    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[.firstnonmetacolumn:ncol(data_clean)]), ]
```

```
## # A tibble: 0 x 79
## # i 79 variables: KUK_ID <chr>, FileName <chr>, FileFormat <fct>,
## #   subcorpus <fct>, SourceID <chr>, DocumentVersion <fct>,
## #   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:ncol(data_clean), ~ scale(.x)))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `across(.firstnonmetacolumn:ncol(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: 63 x 2
##    feat_name                                  p_value
##    <chr>                                        <dbl>
##  1 RuleAbstractNouns                         1.87e- 3
##  2 RuleAnaphoricReferences                   6.60e- 1
##  3 RuleCaseRepetition.max_repetition_count   7.22e- 2
##  4 RuleCaseRepetition.max_repetition_count.v 4.79e- 3
##  5 RuleConfirmationExpressions               9.85e- 2
##  6 RuleDoubleAdpos                           3.12e- 1
##  7 RuleDoubleAdpos.max_allowable_distance    1.90e- 4
##  8 RuleDoubleAdpos.max_allowable_distance.v  3.56e- 6
##  9 RuleInfVerbDistance                       3.55e-15
## 10 RuleInfVerbDistance.max_distance          5.57e- 2
## # i 53 more rows
```

```
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)
```

Correlations

See Levshina (2015: 353–54).

```
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))
```

what unites the low-communality variables we threw out:

- variations have little to do with any other variables in the dataset; there is no factor stemming from the remainder of the feature set to explain them
-

High correlations

```
.hcorrcutoff <- 0.9

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

```
## # A tibble: 22 x 4
##    feat1                        feat2                          cor abs_cor
##    <chr>                        <chr>                        <dbl>   <dbl>
##  1 RuleLongSentences.max_length ari                          0.944   0.944
##  2 RuleLongSentences.max_length gf                           0.922   0.922
##  3 ari                          fkgl                         0.984   0.984
##  4 ari                          gf                           0.978   0.978
##  5 ari                          smog                         0.951   0.951
##  6 ari                          RuleLongSentences.max_length 0.944   0.944
##  7 atl                          cli                          0.960   0.960
##  8 cli                          atl                          0.960   0.960
##  9 fkgl                         ari                          0.984   0.984
## 10 fkgl                         gf                           0.967   0.967
## 11 fkgl                         smog                         0.949   0.949
## 12 gf                           smog                         0.987   0.987
## 13 gf                           ari                          0.978   0.978
## 14 gf                           fkgl                         0.967   0.967
```

```
## 15 gf                          RuleLongSentences.max_length 0.922   0.922
## 16 hpoint                      word_count                   0.957   0.957
## 17 maentropy                   mattr                        0.964   0.964
## 18 mattr                       maentropy                    0.964   0.964
## 19 smog                        gf                           0.987   0.987
## 20 smog                        ari                          0.951   0.951
## 21 smog                        fkgl                         0.949   0.949
## 22 word_count                  hpoint                       0.957   0.957
```

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

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

```
## Compare row 7  and column  34 with corr  0.944
##   Means:  0.4 vs 0.208 so flagging column 7
## Compare row 34  and column  40 with corr  0.978
##   Means:  0.382 vs 0.201 so flagging column 34
## Compare row 40  and column  48 with corr  0.987
##   Means:  0.369 vs 0.193 so flagging column 40
## Compare row 48  and column  38 with corr  0.949
##   Means:  0.349 vs 0.187 so flagging column 48
## Compare row 35  and column  36 with corr  0.96
##   Means:  0.261 vs 0.182 so flagging column 35
## Compare row 50  and column  41 with corr  0.957
##   Means:  0.185 vs 0.179 so flagging column 50
## Compare row 42  and column  45 with corr  0.964
##   Means:  0.175 vs 0.179 so flagging column 45
## All correlations <= 0.9
```

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

```
## [1] "RuleLongSentences.max_length" "ari"
## [3] "gf"                           "smog"
## [5] "atl"                          "word_count"
## [7] "mattr"
```

```
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

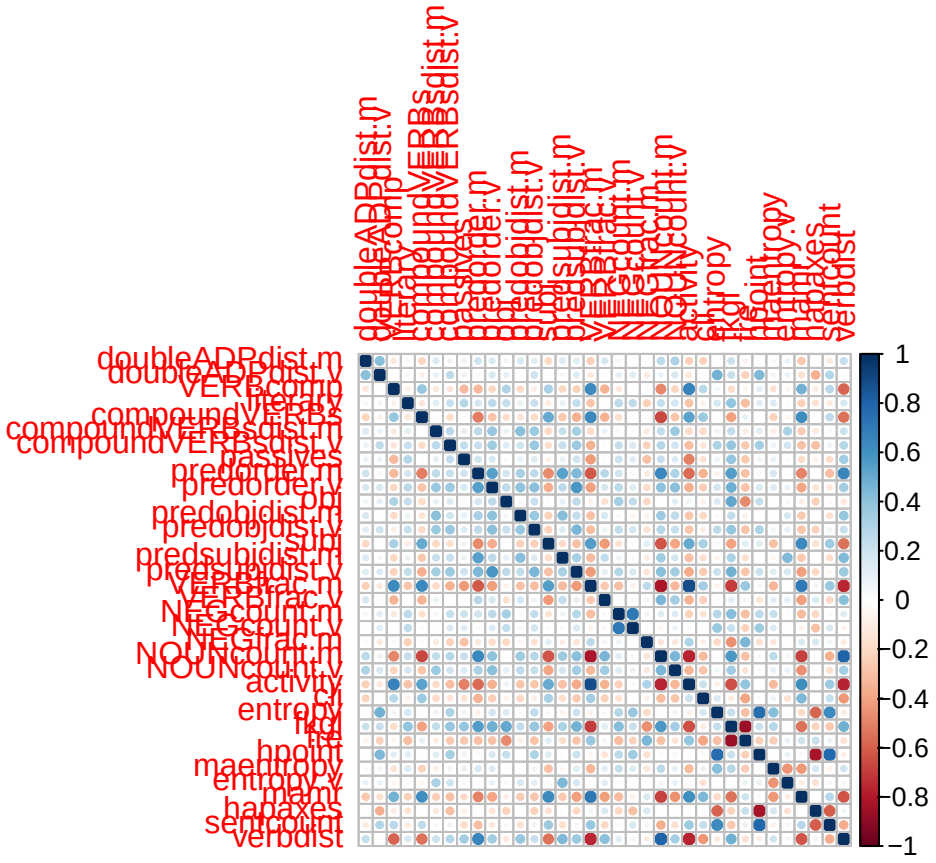
```
# 0.35 instead of 0.3 otherwise the FA bootstrapping would freeze
.lcorrcutoff <- 0.35

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: 9 x 2
##   feat_name                                          p_value
##   <chr>                                                <dbl>
## 1 RuleAbstractNouns                                  0.00187
## 2 RuleCaseRepetition.max_repetition_count.v          0.00479
## 3 RuleRedundantExpressions                           0.0104
## 4 RuleRelativisticExpressions                        0.00205
## 5 RuleTooManyNegations.max_negation_frac.v           0.0365
## 6 RuleTooManyNominalConstructions.max_noun_frac.v 0.00000311
## 7 RuleVerbalNouns                                    0.0000748
## 8 RuleWeakMeaningWords                               0.0386
## 9 GPs                                                0.0138
```

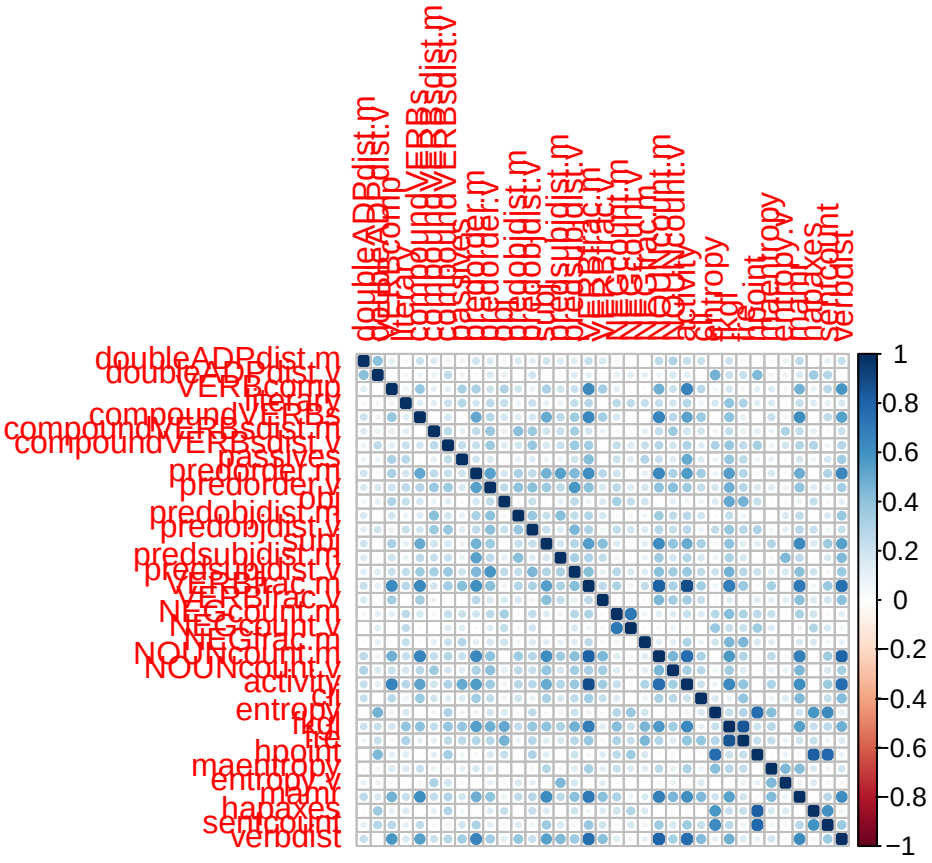
```
data_pure <- data_pureish_striphigh %>%
  select(!any_of(low_correlating_features))

cnames <- map(
  colnames(data_pure),
  function(x) {
    pull(pretty_names %>%
      filter(name_orig == x), name_pretty)
  }
) %>% unlist()

colnames(data_pure) <- cnames
```

Visualisation

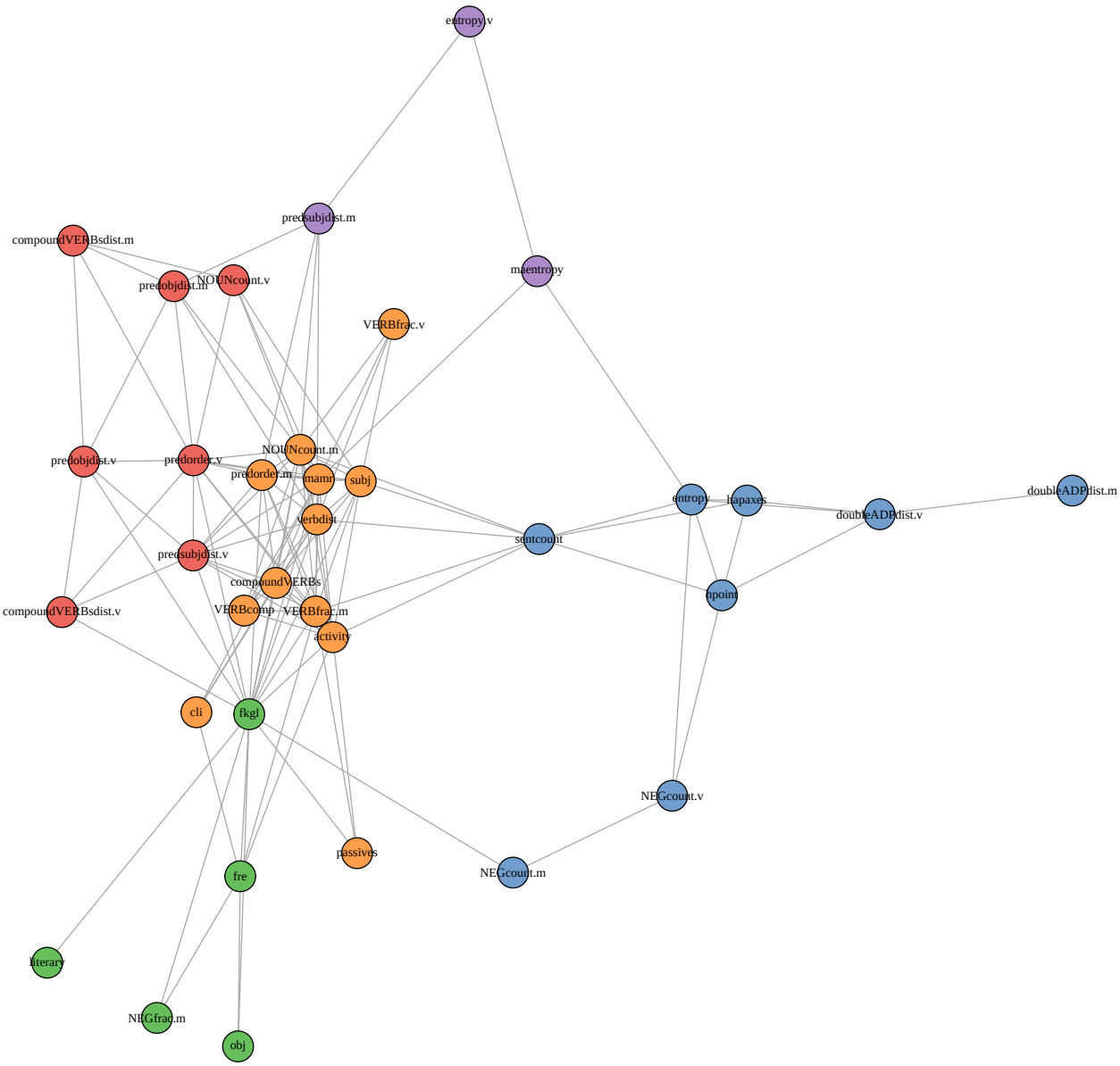
```
corrplot(cor(data_pure))
```




```
corrplot(abs(cor(data_pure)))
```




```
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]))
```

Check for normality

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

```
##          Beta-hat      kappa p-val
## Skewness 1168.858 146886.540     0
## Kurtosis 2987.165    456.508     0
```

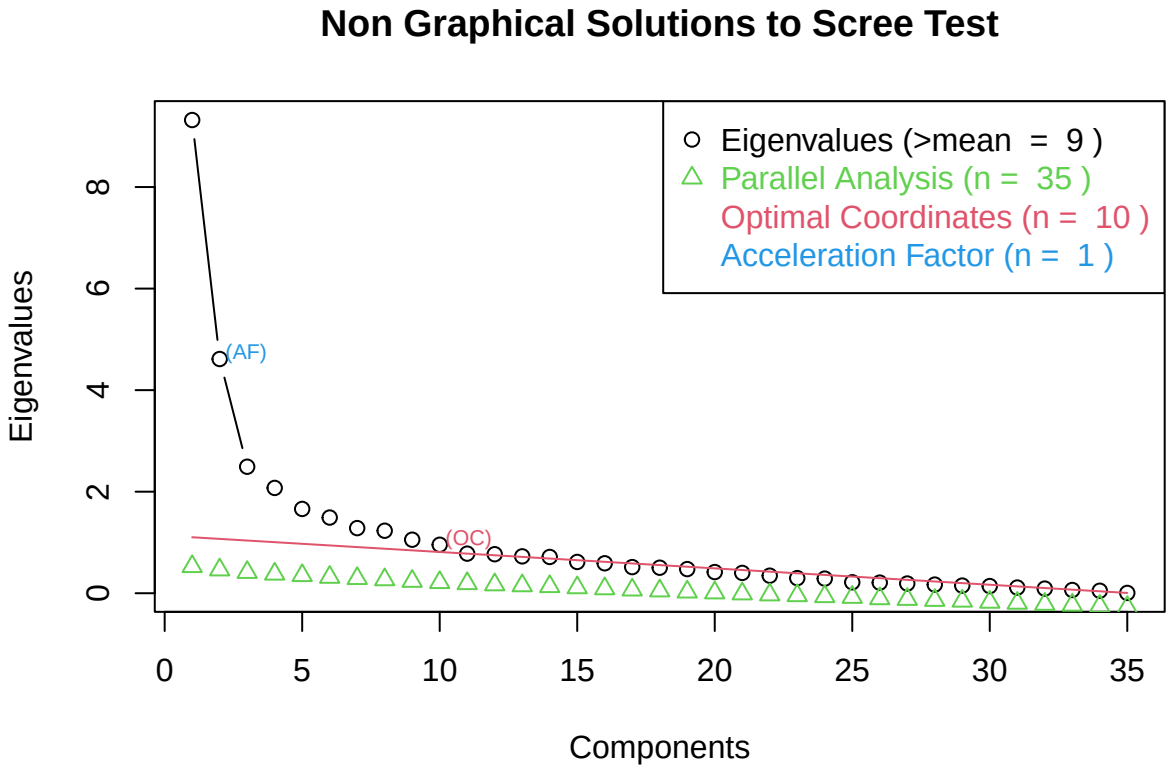
Low (null) p-values show that we can reject the hypothesis that the data would be in a multivariate normal

distribution. I.e. the distribution isn't multivariate normal.

first FA

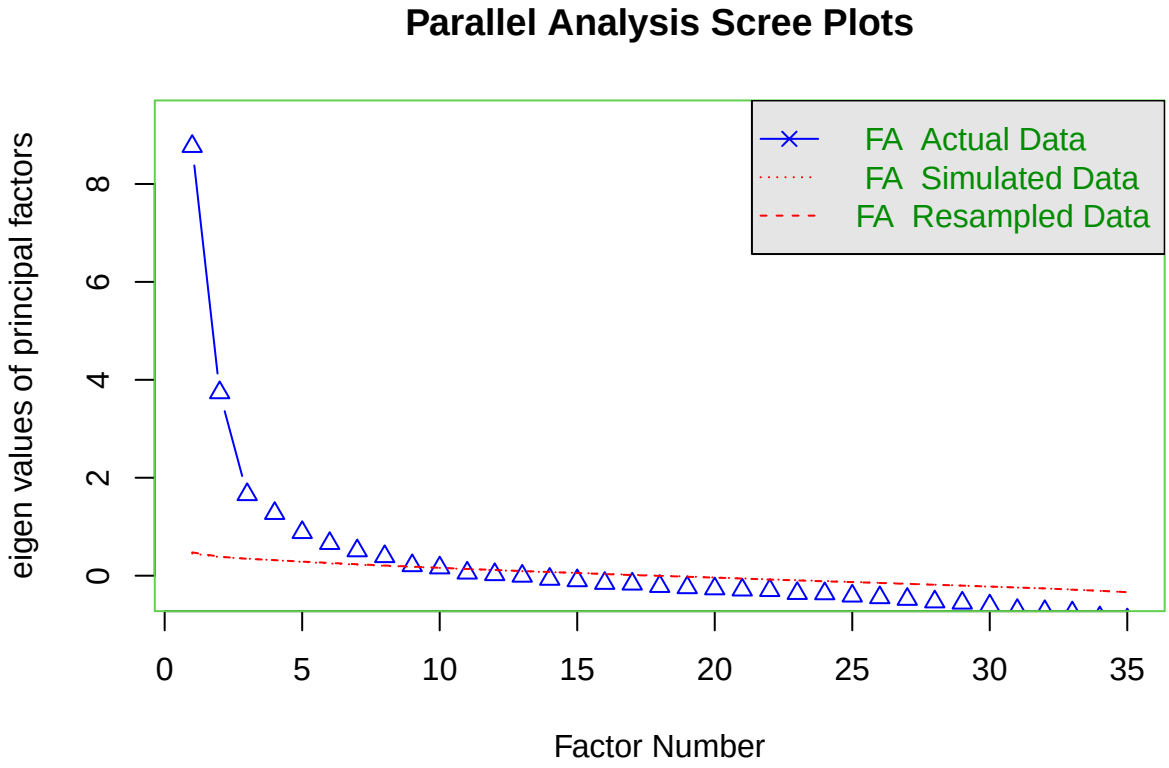
No. of factors

```
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)
```

```
## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs = np.obs, :
## The estimated weights for the factor scores are probably incorrect.  Try a
## different factor score estimation method.
```




```
## Parallel analysis suggests that the number of factors =  9  and the number of components =  NA
```

Model

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

```
# produces ultra-Heywood cases when nfactors = 9
fa_1 <- fa(
  data_scaled,
  nfactors = 9,
  fm = "pa",
  rotate = "promax",
  oblique.scores = TRUE,
  scores = "tenBerge",
  n.iter = 100
)
```

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

```
## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs = np.obs, :
## The estimated weights for the factor scores are probably incorrect.  Try a
## different factor score estimation method.
```

```
## Warning in fac(r = r, nfactors = nfactors, n.obs = n.obs, rotate = rotate, : An
## ultra-Heywood case was detected.  Examine the results carefully
```

```
fa_1
```

```
## Factor Analysis with confidence intervals using method = fa(r = data_scaled, nfactors = 9, n.iter = 
##     scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Factor Analysis using method =  pa
## Call: fa(r = data_scaled, nfactors = 9, n.iter = 100, rotate = "promax",
##     scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
```

```
## Standardized loadings (pattern matrix) based upon correlation matrix
##                         PA1   PA2   PA7   PA4   PA6   PA5   PA3   PA9   PA8   h2
## doubleADPdist.m     -0.08 -0.08 -0.05  0.08 -0.06  0.00  0.17 -0.06  0.66 0.46
## doubleADPdist.v      0.01  0.39 -0.07  0.01 -0.04  0.00  0.06  0.09  0.58 0.52
## VERBcomp             0.64  0.01  0.06  0.50  0.29 -0.11 -0.03  0.07  0.03 0.59
## literary             0.00 -0.03  0.13  0.16 -0.27  0.13 -0.08 -0.03 -0.03 0.24
## compoundVERBs        1.04 -0.15  0.35 -0.28 -0.33  0.05  0.02  0.16 -0.01 0.73
## compoundVERBsdist.m  0.25 -0.05  0.80 -0.09 -0.10 -0.06  0.12 -0.05  0.05 0.47
## compoundVERBsdist.v -0.12  0.26  0.29  0.00 -0.17  0.02  0.07 -0.03 -0.02 0.33
## passives             0.01 -0.08  0.03 -0.22 -0.83  0.09  0.02 -0.08  0.04 0.56
## predorder.m         -0.66 -0.07  0.15  0.20  0.06 -0.06  0.08 -0.01 -0.12 0.61
## predorder.v         -0.08  0.00  0.65  0.15  0.02  0.02 -0.08 -0.06 -0.07 0.55
## obj                  0.14 -0.05  0.01  0.92  0.18  0.12  0.06 -0.05  0.05 0.70
## predobjdist.m       -0.05 -0.11  0.63 -0.10  0.01 -0.06  0.12  0.05 -0.07 0.38
## predobjdist.v        0.02  0.15  0.53  0.05 -0.03  0.05 -0.01  0.01 -0.03 0.37
## subj                 0.57  0.14 -0.18 -0.04 -0.12  0.06  0.22  0.08 -0.06 0.54
## predsubjdist.m      -0.38 -0.05  0.23  0.05  0.15  0.03  0.44  0.20 -0.06 0.50
## predsubjdist.v      -0.21  0.12  0.46  0.15  0.01  0.06  0.01 -0.09 -0.11 0.48
## VERBfrac.m           0.90 -0.06  0.17  0.03  0.34  0.00  0.05  0.06  0.02 0.91
## VERBfrac.v          -0.48 -0.08  0.09 -0.21  0.24  0.05 -0.01  0.06  0.15 0.38
## NEGcount.m          -0.02 -0.08 -0.09  0.18  0.01  0.97  0.03  0.02 -0.01 0.95
## NEGcount.v           0.22  0.08 -0.01  0.06 -0.05  0.74 -0.01  0.05  0.02 0.60
## NEGfrac.m           -0.05 -0.04 -0.11 -0.16  0.40  0.25  0.09 -0.13 -0.05 0.36
## NOUNcount.m         -0.91  0.02  0.02 -0.03 -0.01 -0.13 -0.04  0.03  0.03 0.81
## NOUNcount.v         -0.08 -0.11  0.42  0.03  0.00  0.01 -0.07 -0.09  0.28 0.41
## activity             0.80 -0.01  0.12  0.26  0.47  0.00  0.03 -0.11 -0.05 0.93
## cli                  0.34 -0.01 -0.10 -0.12  0.10  0.03 -0.10  0.76 -0.06 0.70
## entropy              0.01  0.78  0.16 -0.16  0.08  0.09 -0.34  0.16  0.04 0.86
## fkgl                -0.40 -0.06 -0.02  0.52 -0.25  0.07  0.00  0.22  0.10 0.96
## fre                  0.09  0.08  0.10 -0.46  0.17 -0.09  0.01 -0.67 -0.09 1.01
## hpoint              -0.06  0.98 -0.01  0.03  0.00  0.00  0.04 -0.07  0.01 0.94
## maentropy           -0.33  0.04  0.06 -0.12  0.11  0.05 -0.78  0.18 -0.14 0.73
## entropy.v            0.03  0.09  0.26 -0.03  0.11  0.05  0.56  0.01  0.10 0.42
## mamr                 0.70 -0.06 -0.11 -0.01 -0.05 -0.04  0.23  0.14 -0.12 0.73
## hapaxes              0.08 -0.82  0.11 -0.13  0.09  0.01 -0.23  0.12 -0.02 0.74
## sentcount            0.14  0.93  0.01 -0.22  0.25 -0.08 -0.01  0.03 -0.07 0.86
## verbdist            -0.87 -0.02  0.02 -0.20 -0.18 -0.07  0.08 -0.06 -0.06 0.79
##                         u2 com
## doubleADPdist.m      0.5440 1.3
## doubleADPdist.v      0.4751 1.9
## VERBcomp             0.4063 2.5
## literary             0.7583 3.1
## compoundVERBs        0.2653 1.7
## compoundVERBsdist.m  0.5329 1.3
## compoundVERBsdist.v  0.6703 3.1
## passives             0.4365 1.2
## predorder.m          0.3914 1.5
## predorder.v          0.4514 1.2
## obj                  0.2996 1.2
## predobjdist.m        0.6174 1.3
## predobjdist.v        0.6271 1.2
## subj                 0.4622 1.9
## predsubjdist.m       0.5039 3.4
## predsubjdist.v       0.5202 2.1
```

```
## VERBfrac.m           0.0918 1.4
## VERBfrac.v           0.6245 2.4
## NEGcount.m           0.0546 1.1
## NEGcount.v           0.3983 1.2
## NEGfrac.m            0.6398 2.8
## NOUNcount.m          0.1879 1.0
## NOUNcount.v          0.5917 2.2
## activity             0.0727 2.0
## cli                  0.2983 1.6
## entropy              0.1429 1.7
## fkgl                 0.0397 3.0
## fre                 -0.0086 2.1
## hpoint               0.0562 1.0
## maentropy            0.2746 1.7
## entropy.v            0.5816 1.7
## mamr                 0.2703 1.5
## hapaxes              0.2649 1.3
## sentcount            0.1409 1.3
## verbdist             0.2053 1.3
##
##                        PA1  PA2  PA7  PA4  PA6  PA5  PA3  PA9  PA8
## SS loadings           6.75 3.37 2.52 2.04 1.88 1.68 1.49 1.36 1.01
## Proportion Var        0.19 0.10 0.07 0.06 0.05 0.05 0.04 0.04 0.03
## Cumulative Var        0.19 0.29 0.36 0.42 0.47 0.52 0.56 0.60 0.63
## Proportion Explained  0.31 0.15 0.11 0.09 0.08 0.08 0.07 0.06 0.05
## Cumulative Proportion 0.31 0.46 0.57 0.66 0.75 0.82 0.89 0.95 1.00
##
##  With factor correlations of
##       PA1   PA2   PA7   PA4   PA6   PA5   PA3   PA9   PA8
## PA1  1.00  0.10 -0.62 -0.24  0.35 -0.27  0.05 -0.08 -0.31
## PA2  0.10  1.00  0.19  0.31 -0.25  0.29 -0.07  0.22  0.11
## PA7 -0.62  0.19  1.00  0.38 -0.34  0.30  0.06  0.13  0.33
## PA4 -0.24  0.31  0.38  1.00 -0.45  0.24 -0.13  0.30  0.05
## PA6  0.35 -0.25 -0.34 -0.45  1.00 -0.25  0.12 -0.33 -0.03
## PA5 -0.27  0.29  0.30  0.24 -0.25  1.00 -0.19 -0.03  0.11
## PA3  0.05 -0.07  0.06 -0.13  0.12 -0.19  1.00 -0.10 -0.15
## PA9 -0.08  0.22  0.13  0.30 -0.33 -0.03 -0.10  1.00  0.00
## PA8 -0.31  0.11  0.33  0.05 -0.03  0.11 -0.15  0.00  1.00
##
## Mean item complexity =  1.8
## Test of the hypothesis that 9 factors are sufficient.
##
## df null model =  595  with the objective function =  28.74 with Chi Square =  21280.42
## df of  the model are 316  and the objective function was  4.25
##
## 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  754 with the empirical chi square  618.11  with prob <  8.9e-22
## The total n.obs was  754  with Likelihood Chi Square =  3120.31  with prob <  0
##
## Tucker Lewis Index of factoring reliability =  0.743
## RMSEA index =  0.108  and the 90 % confidence intervals are  0.105 0.112
## BIC =  1026.68
```

```
## Fit based upon off diagonal values = 0.99
##  Coefficients and bootstrapped confidence intervals
##                         low   PA1 upper   low   PA2 upper   low   PA7 upper   low
## doubleADPdist.m       -0.26 -0.08  0.04 -0.15 -0.08  0.04 -0.20 -0.05  0.16 -0.05
## doubleADPdist.v       -0.14  0.01  0.14  0.31  0.39  0.51 -0.15 -0.07  0.11 -0.09
## VERBcomp               0.45  0.64  0.78 -0.04  0.01  0.07 -0.02  0.06  0.11  0.29
## literary              -0.13  0.00  0.07 -0.10 -0.03  0.04 -0.02  0.13  0.21  0.04
## compoundVERBs          0.71  1.04  1.23 -0.20 -0.15 -0.07  0.17  0.35  0.43 -0.42
## compoundVERBsdist.m    0.08  0.25  0.36 -0.12 -0.05  0.05  0.54  0.80  0.99 -0.19
## compoundVERBsdist.v   -0.25 -0.12  0.01  0.19  0.26  0.33  0.13  0.29  0.43 -0.10
## passives              -0.15  0.01  0.08 -0.14 -0.08 -0.01 -0.09  0.03  0.11 -0.39
## predorder.m           -0.85 -0.66 -0.44 -0.12 -0.07  0.00  0.03  0.15  0.30  0.05
## predorder.v           -0.25 -0.08  0.06 -0.07  0.00  0.08  0.37  0.65  0.86  0.01
## obj                    0.07  0.14  0.21 -0.11 -0.05 -0.01 -0.06  0.01  0.08  0.52
## predobjdist.m         -0.26 -0.05  0.20 -0.19 -0.11 -0.02  0.34  0.63  0.94 -0.19
## predobjdist.v         -0.11  0.02  0.14  0.04  0.15  0.26  0.29  0.53  0.74 -0.04
## subj                   0.38  0.57  0.77  0.07  0.14  0.20 -0.27 -0.18 -0.07 -0.12
## predsubjdist.m        -0.49 -0.38 -0.21 -0.11 -0.05  0.01  0.07  0.23  0.42 -0.09
## predsubjdist.v        -0.35 -0.21 -0.07  0.04  0.12  0.19  0.23  0.46  0.67  0.06
## VERBfrac.m             0.61  0.90  1.12 -0.09 -0.06 -0.02  0.06  0.17  0.22 -0.01
## VERBfrac.v            -0.62 -0.48 -0.33 -0.15 -0.08  0.00 -0.04  0.09  0.21 -0.29
## NEGcount.m            -0.07 -0.02  0.05 -0.12 -0.08 -0.01 -0.13 -0.09  0.02  0.06
## NEGcount.v             0.13  0.22  0.30  0.02  0.08  0.15 -0.08 -0.01  0.06 -0.02
## NEGfrac.m             -0.12 -0.05  0.09 -0.11 -0.04  0.04 -0.19 -0.11  0.03 -0.28
## NOUNcount.m           -1.09 -0.91 -0.65 -0.03  0.02  0.07 -0.07  0.02  0.12 -0.11
## NOUNcount.v           -0.28 -0.08  0.01 -0.17 -0.11 -0.01  0.22  0.42  0.58 -0.06
## activity               0.57  0.80  0.97 -0.04 -0.01  0.02  0.05  0.12  0.16  0.13
## cli                    0.18  0.34  0.57 -0.08 -0.01  0.07 -0.21 -0.10  0.05 -0.20
## entropy               -0.06  0.01  0.06  0.72  0.78  0.85  0.05  0.16  0.23 -0.25
## fkgl                  -0.50 -0.40 -0.27 -0.09 -0.06 -0.03 -0.07 -0.02  0.05  0.28
## fre                   -0.02  0.09  0.13  0.03  0.08  0.11 -0.03  0.10  0.15 -0.75
## hpoint                -0.10 -0.06  0.00  0.93  0.98  1.03 -0.06 -0.01  0.05 -0.02
## maentropy             -0.45 -0.33 -0.22 -0.03  0.04  0.10 -0.10  0.06  0.14 -0.24
## entropy.v             -0.13  0.03  0.21  0.01  0.09  0.19  0.03  0.26  0.52 -0.12
## mamr                   0.49  0.70  0.92 -0.12 -0.06  0.01 -0.21 -0.11  0.02 -0.09
## hapaxes               -0.01  0.08  0.13 -0.87 -0.82 -0.77  0.00  0.11  0.19 -0.22
## sentcount              0.07  0.14  0.25  0.86  0.93  1.00 -0.04  0.01  0.07 -0.35
## verbdist              -1.05 -0.87 -0.64 -0.05 -0.02  0.01 -0.04  0.02  0.09 -0.32
##                         PA4 upper   low   PA6 upper   low   PA5 upper   low   PA3
## doubleADPdist.m        0.08  0.21 -0.27 -0.06  0.18 -0.17  0.00  0.12 -0.01  0.17
## doubleADPdist.v        0.01  0.11 -0.19 -0.04  0.15 -0.16  0.00  0.12 -0.10  0.06
## VERBcomp               0.50  0.75  0.04  0.29  0.66 -0.19 -0.11 -0.03 -0.08 -0.03
## literary               0.16  0.29 -0.44 -0.27 -0.11  0.01  0.13  0.37 -0.16 -0.08
## compoundVERBs         -0.28 -0.15 -0.54 -0.33 -0.14 -0.03  0.05  0.16 -0.04  0.02
## compoundVERBsdist.m   -0.09  0.03 -0.21 -0.10  0.01 -0.15 -0.06 -0.01 -0.07  0.12
## compoundVERBsdist.v    0.00  0.11 -0.40 -0.17  0.01 -0.07  0.02  0.14 -0.06  0.07
## passives              -0.22 -0.09 -1.30 -0.83 -0.40 -0.03  0.09  0.31 -0.06  0.02
## predorder.m            0.20  0.36 -0.11  0.06  0.25 -0.19 -0.06  0.08 -0.12  0.08
## predorder.v            0.15  0.30 -0.11  0.02  0.16 -0.09  0.02  0.19 -0.18 -0.08
## obj                    0.92  1.35  0.01  0.18  0.44 -0.02  0.12  0.40 -0.01  0.06
## predobjdist.m         -0.10  0.01 -0.16  0.01  0.15 -0.27 -0.06  0.08 -0.17  0.12
## predobjdist.v          0.05  0.17 -0.16 -0.03  0.09 -0.09  0.05  0.20 -0.10 -0.01
## subj                  -0.04  0.04 -0.30 -0.12  0.01 -0.04  0.06  0.16  0.03  0.22
## predsubjdist.m         0.05  0.25 -0.02  0.15  0.30 -0.20  0.03  0.22  0.19  0.44
```

```
## predsubjdist.v       0.15  0.24 -0.16  0.01  0.18 -0.03  0.06  0.19 -0.13  0.01
## VERBfrac.m           0.03  0.10  0.09  0.34  0.67 -0.09  0.00  0.07 -0.01  0.05
## VERBfrac.v          -0.21 -0.11  0.09  0.24  0.37 -0.16  0.05  0.18 -0.20 -0.01
## NEGcount.m           0.18  0.37 -0.11  0.01  0.07  0.66  0.97  1.47 -0.12  0.03
## NEGcount.v           0.06  0.20 -0.17 -0.05  0.04  0.50  0.74  1.26 -0.15 -0.01
## NEGfrac.m           -0.16 -0.04  0.13  0.40  0.71  0.12  0.25  0.40 -0.06  0.09
## NOUNcount.m         -0.03  0.06 -0.13 -0.01  0.05 -0.29 -0.13 -0.05 -0.16 -0.04
## NOUNcount.v          0.03  0.12 -0.09  0.00  0.14 -0.09  0.01  0.12 -0.18 -0.07
## activity             0.26  0.41  0.10  0.47  0.99 -0.04  0.00  0.07  0.00  0.03
## cli                 -0.12  0.03 -0.19  0.10  0.25 -0.26  0.03  0.17 -0.43 -0.10
## entropy             -0.16 -0.07  0.00  0.08  0.19  0.02  0.09  0.22 -0.73 -0.34
## fkgl                 0.52  0.79 -0.53 -0.25 -0.04  0.02  0.07  0.16 -0.06  0.00
## fre                 -0.46 -0.23 -0.08  0.17  0.58 -0.16 -0.09  0.03 -0.05  0.01
## hpoint               0.03  0.06 -0.04  0.00  0.06 -0.04  0.00  0.08 -0.02  0.04
## maentropy           -0.12  0.00 -0.03  0.11  0.33 -0.02  0.05  0.24 -1.59 -0.78
## entropy.v           -0.03  0.07 -0.04  0.11  0.22 -0.14  0.05  0.18  0.21  0.56
## mamr                -0.01  0.09 -0.19 -0.05  0.04 -0.18 -0.04  0.05  0.07  0.23
## hapaxes             -0.13 -0.03  0.00  0.09  0.19 -0.07  0.01  0.08 -0.52 -0.23
## sentcount           -0.22 -0.10  0.08  0.25  0.44 -0.20 -0.08 -0.01 -0.09 -0.01
## verbdist            -0.20 -0.10 -0.48 -0.18  0.00 -0.21 -0.07  0.03 -0.02  0.08
##                     upper   low   PA9 upper   low   PA8 upper
## doubleADPdist.m      0.37 -0.46 -0.06  0.26  0.25  0.66  1.10
## doubleADPdist.v      0.16 -0.04  0.09  0.26  0.27  0.58  0.89
## VERBcomp             0.07 -0.08  0.07  0.37 -0.07  0.03  0.13
## literary             0.04 -0.20 -0.03  0.28 -0.14 -0.03  0.11
## compoundVERBs        0.15 -0.15  0.16  0.67 -0.20 -0.01  0.12
## compoundVERBsdist.m  0.42 -0.21 -0.05  0.09 -0.12  0.05  0.27
## compoundVERBsdist.v  0.24 -0.17 -0.03  0.14 -0.14 -0.02  0.13
## passives             0.19 -0.29 -0.08  0.28 -0.17  0.04  0.21
## predorder.m          0.31 -0.46 -0.01  0.32 -0.37 -0.12  0.06
## predorder.v          0.09 -0.21 -0.06  0.08 -0.22 -0.07  0.09
## obj                  0.22 -0.18 -0.05  0.18 -0.05  0.05  0.19
## predobjdist.m        0.38 -0.25  0.05  0.29 -0.28 -0.07  0.13
## predobjdist.v        0.11 -0.23  0.01  0.27 -0.17 -0.03  0.15
## subj                 0.44 -0.09  0.08  0.27 -0.27 -0.06  0.09
## predsubjdist.m       0.77 -0.04  0.20  0.50 -0.38 -0.06  0.23
## predsubjdist.v       0.23 -0.25 -0.09  0.04 -0.25 -0.11  0.04
## VERBfrac.m           0.13 -0.03  0.06  0.17 -0.07  0.02  0.09
## VERBfrac.v           0.09 -0.23  0.06  0.25  0.00  0.15  0.37
## NEGcount.m           0.07 -0.36  0.02  0.22 -0.15 -0.01  0.07
## NEGcount.v           0.08 -0.15  0.05  0.20 -0.10  0.02  0.11
## NEGfrac.m            0.17 -0.84 -0.13  0.23 -0.29 -0.05  0.13
## NOUNcount.m          0.03 -0.18  0.03  0.16 -0.08  0.03  0.20
## NOUNcount.v          0.05 -0.27 -0.09  0.07  0.03  0.28  0.60
## activity             0.13 -0.24 -0.11 -0.02 -0.13 -0.05  0.02
## cli                  0.00  0.31  0.76  1.50 -0.22 -0.06  0.09
## entropy             -0.09 -0.03  0.16  0.55 -0.08  0.04  0.22
## fkgl                 0.04 -0.03  0.22  0.69  0.00  0.10  0.24
## fre                  0.21 -1.51 -0.67 -0.17 -0.23 -0.09  0.03
## hpoint               0.13 -0.20 -0.07  0.02 -0.08  0.01  0.16
## maentropy           -0.26 -0.09  0.18  0.78 -0.35 -0.14  0.04
## entropy.v            1.11 -0.39  0.01  0.26 -0.12  0.10  0.42
## mamr                 0.47 -0.03  0.14  0.33 -0.36 -0.12  0.05
## hapaxes             -0.05 -0.02  0.12  0.35 -0.17 -0.02  0.09
```

```
## sentcount            0.05 -0.15  0.03  0.15 -0.14 -0.07  0.04
## verbdist             0.18 -0.31 -0.06  0.06 -0.14 -0.06  0.03
##
##  Interfactor correlations and bootstrapped confidence intervals
##         lower estimate upper
## PA1-PA2 -0.17   0.0967  0.32
## PA1-PA7 -0.91  -0.6207  0.50
## PA1-PA4 -0.99  -0.2376  0.24
## PA1-PA6 -0.73   0.3499  0.37
## PA1-PA5 -0.59  -0.2707  0.33
## PA1-PA3 -0.46   0.0454  0.37
## PA1-PA9 -0.45  -0.0767  0.32
## PA1-PA8 -0.52  -0.3059  0.21
## PA2-PA7 -0.30   0.1877  0.57
## PA2-PA4 -0.17   0.3053  0.52
## PA2-PA6 -0.28  -0.2468  0.61
## PA2-PA5 -0.19   0.2910  0.47
## PA2-PA3 -0.24  -0.0695  0.35
## PA2-PA9 -0.17   0.2241  0.30
## PA2-PA8 -0.15   0.1123  0.29
## PA7-PA4 -0.64   0.3833  0.86
## PA7-PA6 -0.66  -0.3442  0.80
## PA7-PA5 -0.54   0.2951  0.55
## PA7-PA3 -0.41   0.0601  0.43
## PA7-PA9 -0.41   0.1252  0.36
## PA7-PA8 -0.35   0.3267  0.41
## PA4-PA6 -0.49  -0.4451  0.79
## PA4-PA5 -0.38   0.2418  0.60
## PA4-PA3 -0.35  -0.1289  0.44
## PA4-PA9 -0.41   0.2967  0.44
## PA4-PA8 -0.34   0.0521  0.45
## PA6-PA5 -0.37  -0.2453  0.38
## PA6-PA3 -0.40   0.1152  0.35
## PA6-PA9 -0.42  -0.3301  0.37
## PA6-PA8 -0.35  -0.0263  0.35
## PA5-PA3 -0.33  -0.1919  0.33
## PA5-PA9 -0.36  -0.0278  0.34
## PA5-PA8 -0.27   0.1108  0.36
## PA3-PA9 -0.37  -0.0975  0.32
## PA3-PA8 -0.28  -0.1472  0.33
## PA9-PA8 -0.22  -0.0015  0.32
```

Healthiness diagnostics

```
fa_1$loadings[] %>%
  as_tibble() %>%
  mutate(feat = cnames) %>%
  select(feat, everything()) %>%
  pivot_longer(!feat) %>%
  mutate(value = abs(value)) %>%
  group_by(feat) %>%
  summarize(maxload = max(value)) %>%
  arrange(maxload)
```

```
## # A tibble: 35 x 2
##    feat                maxload
##    <chr>                 <dbl>
##  1 literary              0.267
##  2 compoundVERBsdist.v   0.292
##  3 NEGfrac.m             0.400
##  4 NOUNcount.v           0.424
##  5 predsubjdist.m        0.437
##  6 predsubjdist.v        0.463
##  7 VERBfrac.v            0.479
##  8 fkgl                  0.516
##  9 predobjdist.v         0.525
## 10 entropy.v             0.561
## # i 25 more rows
```

```
fa_1$communality %>% sort()
```

```
##            literary compoundVERBsdist.v           NEGfrac.m       predobjdist.v
##           0.2416525           0.3296794           0.3601689           0.3729226
##          VERBfrac.v       predobjdist.m         NOUNcount.v           entropy.v
##           0.3755086           0.3826177           0.4082917           0.4184301
##     doubleADPdist.m compoundVERBsdist.m      predsubjdist.v      predsubjdist.m
##           0.4560091           0.4671012           0.4798399           0.4961259
##     doubleADPdist.v                subj         predorder.v            passives
##           0.5248710           0.5377723           0.5485923           0.5634874
##            VERBcomp          NEGcount.v         predorder.m                 obj
##           0.5936647           0.6016981           0.6085826           0.7004213
##                 cli           maentropy                mamr       compoundVERBs
##           0.7017128           0.7254411           0.7296938           0.7347473
##             hapaxes            verbdist         NOUNcount.m             entropy
##           0.7350817           0.7946721           0.8120502           0.8571486
##           sentcount          VERBfrac.m            activity              hpoint
##           0.8590817           0.9081530           0.9273428           0.9438129
##          NEGcount.m                fkgl                 fre
##           0.9454264           0.9603051           1.0085624
```

```
fa_1$communality[fa_1$communality < 0.5] %>% names()
```

```
##  [1] "doubleADPdist.m"     "literary"            "compoundVERBsdist.m"
##  [4] "compoundVERBsdist.v" "predobjdist.m"       "predobjdist.v"
##  [7] "predsubjdist.m"      "predsubjdist.v"      "VERBfrac.v"
## [10] "NEGfrac.m"           "NOUNcount.v"         "entropy.v"
```

```
fa_1$complexity %>% sort()
```

```
##              hpoint         NOUNcount.m          NEGcount.m                 obj
##            1.021970            1.049967            1.107250            1.188772
##         predorder.v       predobjdist.v            passives          NEGcount.v
##            1.212926            1.217932            1.224465            1.226073
##            verbdist       predobjdist.m     doubleADPdist.m             hapaxes
##            1.257344            1.271641            1.280541            1.333635
##           sentcount compoundVERBsdist.m          VERBfrac.m                mamr
##            1.343466            1.349938            1.381732            1.456405
##         predorder.m                 cli           entropy.v           maentropy
##            1.477622            1.599286            1.689031            1.690295
##             entropy       compoundVERBs     doubleADPdist.v                subj
```

```
##           1.724179           1.730186           1.880323           1.896206
##           activity      predsubjdist.v                fre        NOUNcount.v
##           2.005256           2.074407           2.141251           2.172268
##          VERBfrac.v           VERBcomp          NEGfrac.m               fkgl
##           2.419905           2.484551           2.775734           2.984126
##           literary compoundVERBsdist.v      predsubjdist.m
##           3.093669           3.146079           3.403237
```

```
fa_1$complexity[fa_1$complexity > 2] %>% names()
```

```
##  [1] "VERBcomp"            "literary"            "compoundVERBsdist.v"
##  [4] "predsubjdist.m"      "predsubjdist.v"      "VERBfrac.v"
##  [7] "NEGfrac.m"           "NOUNcount.v"         "activity"
## [10] "fkgl"                "fre"
```

Feature engineering

```
data_engineered_1 <- data_scaled %>%
  # remove low-communality variables
  select(!c(
    doubleADPdist.m,
    doubleADPdist.v,
    literary,
    compoundVERBsdist.m,
    compoundVERBsdist.v,
    predobjdist.m,
    predobjdist.v,
    predsubjdist.v,
    VERBfrac.v,
    NEGfrac.m,
    NOUNcount.v,
    entropy.v
  )) %>%
  # remove confound variables
  select(!c(cli, fkgl, fre))

det(cor(data_engineered_1))
```

```
## [1] 1.282125e-07
```

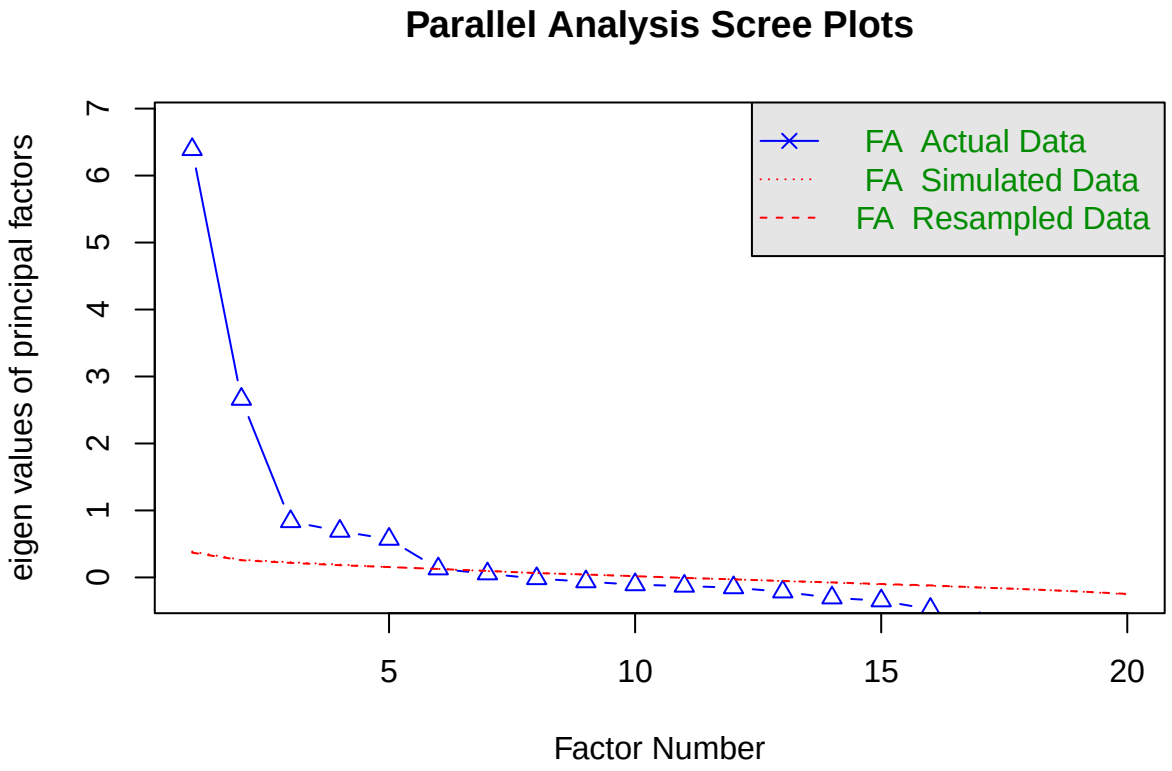
```
KMO(data_engineered_1)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = data_engineered_1)
## Overall MSA =  0.83
## MSA for each item =
##       VERBcomp  compoundVERBs       passives    predorder.m    predorder.v
##           0.86           0.90           0.77           0.85           0.83
##            obj           subj predsubjdist.m     VERBfrac.m     NEGcount.m
##           0.56           0.93           0.80           0.88           0.72
##     NEGcount.v    NOUNcount.m       activity        entropy         hpoint
##           0.67           0.92           0.89           0.69           0.70
##      maentropy           mamr        hapaxes      sentcount       verbdist
##           0.60           0.91           0.77           0.74           0.92
```

second FA

No. of vectors

```
fa.parallel(data_engineered_1, fm = "pa", fa = "fa", n.iter = 20)
```




```
## Parallel analysis suggests that the number of factors =  5  and the number of components =  NA
```

Model

```
fa_2 <- fa(
  data_engineered_1,
  nfactors = 5,
  fm = "pa",
  rotate = "promax",
  oblique.scores = TRUE,
  scores = "tenBerge",
  n.iter = 100
)
fa_2
```

```
## Factor Analysis with confidence intervals using method = fa(r = data_engineered_1, nfactors = 5, n.i
##     scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Factor Analysis using method =  pa
## Call: fa(r = data_engineered_1, nfactors = 5, n.iter = 100, rotate = "promax",
##     scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##                   PA1   PA2   PA4   PA3   PA5   h2    u2 com
## VERBcomp         0.23  0.05  0.61  0.05 -0.04 0.56 0.437 1.3
## compoundVERBs    0.75  0.00 -0.12  0.09 -0.17 0.55 0.454 1.2
```

```
## passives        0.03  0.01 -0.60  0.23 -0.12 0.35 0.653 1.4
## predorder.m    -0.85 -0.03  0.02  0.00 -0.16 0.69 0.315 1.1
## predorder.v    -0.54  0.10  0.05  0.16 -0.02 0.35 0.649 1.3
## obj            -0.31  0.00  0.45  0.41 -0.05 0.46 0.543 2.8
## subj            0.61  0.14 -0.07  0.05 -0.28 0.52 0.481 1.6
## predsubjdist.m -0.54  0.02 -0.02 -0.04 -0.28 0.30 0.696 1.5
## VERBfrac.m      0.64 -0.04  0.42 -0.07 -0.10 0.88 0.116 1.8
## NEGcount.m      0.03 -0.10 -0.16  0.89  0.13 0.76 0.241 1.1
## NEGcount.v      0.26  0.05 -0.18  0.79  0.11 0.62 0.379 1.4
## NOUNcount.m    -0.82  0.04 -0.16 -0.17  0.10 0.81 0.193 1.2
## activity        0.49 -0.05  0.61 -0.02 -0.07 0.89 0.109 2.0
## entropy         0.03  0.76  0.03  0.10  0.46 0.86 0.144 1.7
## hpoint         -0.10  0.98 -0.03  0.03 -0.03 0.96 0.037 1.0
## maentropy      -0.09 -0.02  0.06  0.12  0.71 0.54 0.463 1.1
## mamr            0.65 -0.03  0.03 -0.03 -0.39 0.72 0.282 1.7
## hapaxes         0.14 -0.83  0.07 -0.04  0.25 0.75 0.255 1.3
## sentcount       0.22  0.87  0.10 -0.22  0.03 0.82 0.185 1.3
## verbdist       -0.69 -0.01 -0.39 -0.14 -0.06 0.79 0.211 1.7
##
##                        PA1  PA2  PA4  PA3  PA5
## SS loadings           5.09 3.00 2.04 1.72 1.30
## Proportion Var        0.25 0.15 0.10 0.09 0.07
## Cumulative Var        0.25 0.40 0.51 0.59 0.66
## Proportion Explained  0.39 0.23 0.16 0.13 0.10
## Cumulative Proportion 0.39 0.61 0.77 0.90 1.00
##
##  With factor correlations of
##       PA1  PA2   PA4   PA3   PA5
## PA1  1.00 0.11  0.39 -0.23 -0.21
## PA2  0.11 1.00  0.14  0.37  0.00
## PA4  0.39 0.14  1.00  0.08 -0.32
## PA3 -0.23 0.37  0.08  1.00  0.00
## PA5 -0.21 0.00 -0.32  0.00  1.00
##
## Mean item complexity =  1.5
## Test of the hypothesis that 5 factors are sufficient.
##
## df null model =  190  with the objective function =  15.87 with Chi Square =  11830.77
## df of  the model are 100  and the objective function was  1.88
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.05
##
## The harmonic n.obs is  754 with the empirical chi square  335.06  with prob <  3.9e-27
## The total n.obs was  754  with Likelihood Chi Square =  1393.99  with prob <  7.3e-227
##
## Tucker Lewis Index of factoring reliability =  0.788
## RMSEA index =  0.131  and the 90 % confidence intervals are  0.125 0.137
## BIC =  731.45
## Fit based upon off diagonal values = 0.99
## Measures of factor score adequacy
##                                                 PA1  PA2  PA4  PA3 PA5
## Correlation of (regression) scores with factors   0.97 0.99 0.93 0.93 0.9
## Multiple R square of scores with factors          0.94 0.98 0.87 0.86 0.8
```

```
## Minimum correlation of possible factor scores     0.87 0.96 0.74 0.72 0.6
##
##  Coefficients and bootstrapped confidence intervals
##                   low   PA1 upper   low   PA2 upper   low   PA4 upper   low
## VERBcomp         0.13  0.23  0.33 -0.01  0.05  0.11  0.47  0.61  0.75 -0.01
## compoundVERBs    0.52  0.75  0.93 -0.06  0.00  0.05 -0.23 -0.12 -0.01  0.02
## passives        -0.09  0.03  0.15 -0.06  0.01  0.08 -0.77 -0.60 -0.48  0.13
## predorder.m     -0.94 -0.85 -0.69 -0.09 -0.03  0.02 -0.09  0.02  0.11 -0.07
## predorder.v     -0.65 -0.54 -0.40  0.03  0.10  0.17 -0.05  0.05  0.15  0.08
## obj             -0.40 -0.31 -0.17 -0.05  0.00  0.05  0.33  0.45  0.56  0.32
## subj             0.49  0.61  0.71  0.08  0.14  0.20 -0.15 -0.07  0.01  0.00
## predsubjdist.m -0.63 -0.54 -0.39 -0.03  0.02  0.05 -0.15 -0.02  0.11 -0.11
## VERBfrac.m       0.48  0.64  0.78 -0.07 -0.04  0.00  0.31  0.42  0.55 -0.13
## NEGcount.m      -0.05  0.03  0.09 -0.13 -0.10 -0.05 -0.23 -0.16 -0.08  0.82
## NEGcount.v       0.16  0.26  0.31  0.01  0.05  0.10 -0.25 -0.18 -0.09  0.73
## NOUNcount.m     -0.97 -0.82 -0.61  0.00  0.04  0.07 -0.25 -0.16 -0.09 -0.23
## activity         0.38  0.49  0.59 -0.08 -0.05 -0.01  0.50  0.61  0.75 -0.06
## entropy         -0.04  0.03  0.09  0.72  0.76  0.81 -0.05  0.03  0.10  0.05
## hpoint          -0.13 -0.10 -0.05  0.95  0.98  1.00 -0.07 -0.03  0.01  0.01
## maentropy       -0.19 -0.09  0.01 -0.07 -0.02  0.03 -0.05  0.06  0.15  0.07
## mamr             0.49  0.65  0.79 -0.07 -0.03  0.02 -0.05  0.03  0.10 -0.09
## hapaxes          0.07  0.14  0.20 -0.86 -0.83 -0.79 -0.01  0.07  0.15 -0.09
## sentcount        0.13  0.22  0.30  0.83  0.87  0.92  0.04  0.10  0.16 -0.29
## verbdist        -0.80 -0.69 -0.54 -0.04 -0.01  0.02 -0.55 -0.39 -0.29 -0.23
##                   PA3 upper   low   PA5 upper
## VERBcomp         0.05  0.12 -0.13 -0.04  0.04
## compoundVERBs    0.09  0.18 -0.43 -0.17  0.02
## passives         0.23  0.35 -0.31 -0.12  0.01
## predorder.m      0.00  0.09 -0.48 -0.16  0.11
## predorder.v      0.16  0.25 -0.12 -0.02  0.07
## obj              0.41  0.52 -0.14 -0.05  0.04
## subj             0.05  0.11 -0.48 -0.28 -0.15
## predsubjdist.m -0.04  0.05 -0.58 -0.28 -0.04
## VERBfrac.m      -0.07 -0.01 -0.24 -0.10  0.00
## NEGcount.m       0.89  0.97  0.04  0.13  0.28
## NEGcount.v       0.79  0.85  0.02  0.11  0.24
## NOUNcount.m     -0.17 -0.12  0.00  0.10  0.25
## activity        -0.02  0.03 -0.13 -0.07 -0.02
## entropy          0.10  0.16  0.38  0.46  0.59
## hpoint           0.03  0.06 -0.08 -0.03  0.04
## maentropy        0.12  0.20  0.63  0.71  0.92
## mamr            -0.03  0.02 -0.72 -0.39 -0.19
## hapaxes         -0.04  0.01  0.16  0.25  0.35
## sentcount       -0.22 -0.15 -0.03  0.03  0.10
## verbdist        -0.14 -0.05 -0.13 -0.06  0.01
##
##  Interfactor correlations and bootstrapped confidence intervals
##         lower estimate upper
## PA1-PA2 -0.33   0.1079  0.37
## PA1-PA4 -0.72   0.3904  0.76
## PA1-PA3 -0.76  -0.2280  0.61
## PA1-PA5 -0.41  -0.2124  0.30
## PA2-PA4 -0.15   0.1382  0.43
## PA2-PA3 -0.23   0.3711  0.62
```

```
## PA2-PA5 -0.23  -0.0046  0.44
## PA4-PA3 -0.37   0.0840  0.36
## PA4-PA5 -0.42  -0.3245  0.44
## PA3-PA5 -0.22   0.0013  0.22
```

Healthiness diagnostics

```
fa_2$loadings[] %>%
  as_tibble() %>%
  mutate(feat = colnames(data_engineered_1)) %>%
  select(feat, everything()) %>%
  pivot_longer(!feat) %>%
  mutate(value = abs(value)) %>%
  group_by(feat) %>%
  summarize(maxload = max(value)) %>%
  arrange(maxload)
```

```
## # A tibble: 20 x 2
##    feat           maxload
##    <chr>            <dbl>
##  1 obj              0.447
##  2 predsubjdist.m   0.544
##  3 predorder.v      0.544
##  4 passives         0.603
##  5 VERBcomp         0.606
##  6 subj             0.613
##  7 activity         0.614
##  8 VERBfrac.m       0.644
##  9 mamr             0.650
## 10 verbdist         0.687
## 11 maentropy        0.714
## 12 compoundVERBs    0.746
## 13 entropy          0.764
## 14 NEGcount.v       0.794
## 15 NOUNcount.m      0.817
## 16 hapaxes          0.829
## 17 predorder.m      0.850
## 18 sentcount        0.870
## 19 NEGcount.m       0.888
## 20 hpoint           0.976
```

```
fa_2$communality %>% sort()
```

```
## predsubjdist.m       passives    predorder.v           obj          subj
##      0.3037867      0.3468791      0.3505151      0.4574124      0.5188582
##      maentropy  compoundVERBs       VERBcomp     NEGcount.v    predorder.m
##      0.5372362      0.5461787      0.5627326      0.6208769      0.6852554
##           mamr        hapaxes     NEGcount.m       verbdist    NOUNcount.m
##      0.7184793      0.7454838      0.7586160      0.7893884      0.8071030
##      sentcount        entropy     VERBfrac.m       activity         hpoint
##      0.8150467      0.8557144      0.8841659      0.8906305      0.9625622
```

```
fa_2$communality[fa_2$communality < 0.5] %>% names()
```

```
## [1] "passives"       "predorder.v"    "obj"            "predsubjdist.m"
```

```
fa_2$complexity %>% sort()
```

```
##         hpoint    predorder.m      maentropy     NEGcount.m  compoundVERBs
##       1.026114       1.070268       1.109904       1.141679       1.197415
##    NOUNcount.m    predorder.v        hapaxes      sentcount       VERBcomp
##       1.203861       1.263880       1.270866       1.295166       1.310073
##     NEGcount.v       passives predsubjdist.m           subj           mamr
##       1.369062       1.393392       1.515239       1.582351       1.661518
##        entropy       verbdist     VERBfrac.m       activity            obj
##       1.688315       1.704079       1.823346       1.950707       2.812552
```

```
fa_2$complexity[fa_2$complexity > 2] %>% names()
```

```
## [1] "obj"
```

Feature engineering

```
data_engineered_2 <- data_engineered_1 %>%
  # remove low-communality features
  select(!c(
    passives,
    predorder.v,
    obj,
    predsubjdist.m
  ))

det(cor(data_engineered_2))
```

```
## [1] 1.289469e-06
```

```
KMO(data_engineered_2)
```

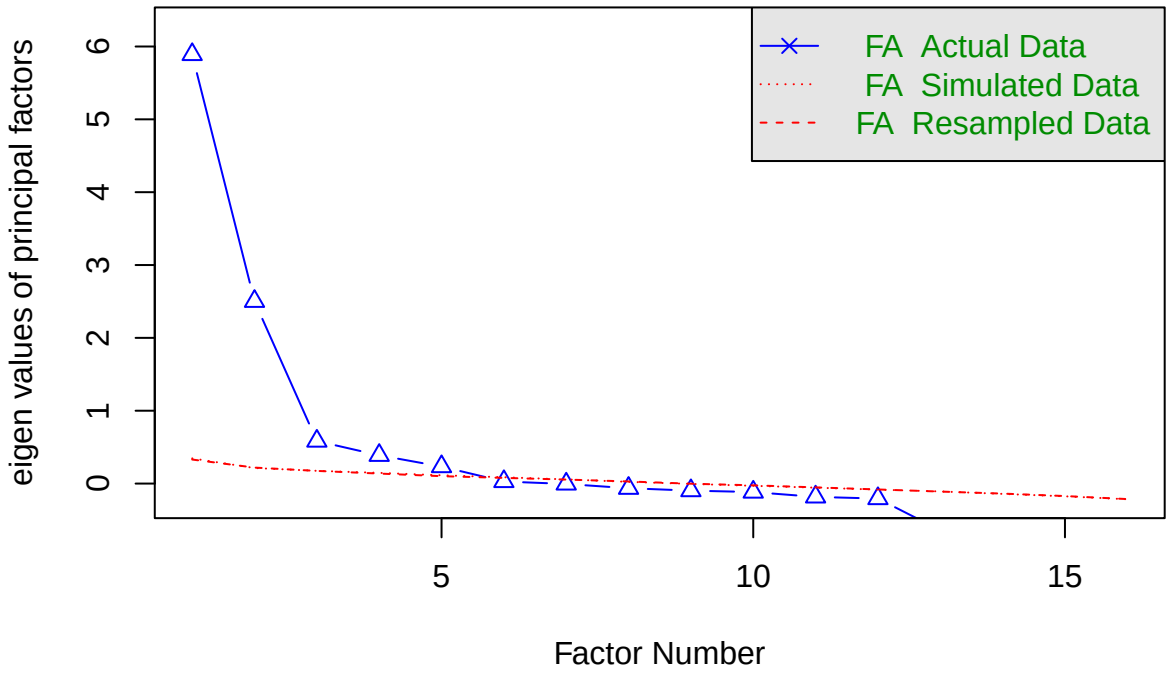
```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = data_engineered_2)
## Overall MSA =  0.84
## MSA for each item =
##      VERBcomp compoundVERBs   predorder.m          subj    VERBfrac.m
##          0.84          0.94          0.94          0.94          0.86
##    NEGcount.m    NEGcount.v   NOUNcount.m      activity       entropy
##          0.66          0.64          0.91          0.88          0.72
##        hpoint     maentropy          mamr       hapaxes     sentcount
##          0.70          0.65          0.90          0.77          0.77
##      verbdist
##          0.90
```

```
final_collist <- data_engineered_2 %>% colnames()
```

Final FA

No. of vectors

```
fa.parallel(data_engineered_2, fm = "pa", fa = "fa", n.iter = 20)
```




```
## Parallel analysis suggests that the number of factors =  5  and the number of components =  NA
```

Model

```
fa_res <- fa(
  data_engineered_2,
  nfactors = 5,
  fm = "pa",
  rotate = "promax",
  oblique.scores = TRUE,
  scores = "tenBerge",
  n.iter = 100
)
fa_res
```

```
## Factor Analysis with confidence intervals using method = fa(r = data_engineered_2, nfactors = 5, n.i
##     scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Factor Analysis using method =  pa
## Call: fa(r = data_engineered_2, nfactors = 5, n.iter = 100, rotate = "promax",
##     scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##                 PA1   PA2   PA5   PA3   PA4   h2    u2 com
## VERBcomp       0.16  0.09  0.60  0.01 -0.01 0.52 0.482 1.2
## compoundVERBs  0.79 -0.05 -0.08  0.02  0.00 0.53 0.465 1.0
## predorder.m   -0.76  0.02  0.02  0.03 -0.12 0.52 0.482 1.1
## subj           0.75  0.11 -0.16  0.00 -0.14 0.54 0.459 1.2
## VERBfrac.m     0.61 -0.06  0.43 -0.06 -0.03 0.90 0.096 1.9
## NEGcount.m    -0.11 -0.05  0.04  0.91  0.00 0.83 0.170 1.0
## NEGcount.v     0.17  0.07 -0.03  0.80  0.02 0.68 0.322 1.1
## NOUNcount.m   -0.89  0.07 -0.09 -0.10 -0.02 0.84 0.165 1.1
```

```
## activity       0.39 -0.03  0.65  0.00 -0.06 0.91 0.095 1.7
## entropy        0.10  0.71 -0.06  0.01  0.55 0.95 0.055 1.9
## hpoint        -0.13  0.99  0.04  0.06 -0.05 0.96 0.040 1.1
## maentropy     -0.08 -0.11 -0.03  0.01  0.77 0.64 0.358 1.1
## mamr           0.74 -0.04 -0.02 -0.05 -0.26 0.71 0.287 1.3
## hapaxes        0.18 -0.88 -0.01 -0.09  0.29 0.77 0.229 1.3
## sentcount      0.22  0.80  0.09 -0.15  0.05 0.77 0.232 1.3
## verbdist      -0.69  0.00 -0.29 -0.07 -0.10 0.75 0.247 1.4
##
##                        PA1  PA2  PA5  PA3  PA4
## SS loadings           4.67 2.95 1.53 1.52 1.15
## Proportion Var        0.29 0.18 0.10 0.10 0.07
## Cumulative Var        0.29 0.48 0.57 0.67 0.74
## Proportion Explained  0.39 0.25 0.13 0.13 0.10
## Cumulative Proportion 0.39 0.64 0.77 0.90 1.00
##
##  With factor correlations of
##       PA1  PA2   PA5   PA3   PA4
## PA1  1.00 0.18  0.60 -0.17 -0.26
## PA2  0.18 1.00  0.07  0.29  0.16
## PA5  0.60 0.07  1.00 -0.17 -0.15
## PA3 -0.17 0.29 -0.17  1.00  0.28
## PA4 -0.26 0.16 -0.15  0.28  1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 5 factors are sufficient.
##
## df null model =  120  with the objective function =  13.56 with Chi Square =  10128.02
## df of  the model are 50  and the objective function was  0.76
##
## The root mean square of the residuals (RMSR) is  0.02
## The df corrected root mean square of the residuals is  0.03
##
## The harmonic n.obs is  754 with the empirical chi square  60.5  with prob <  0.15
## The total n.obs was  754  with Likelihood Chi Square =  562.84  with prob <  6.5e-88
##
## Tucker Lewis Index of factoring reliability =  0.876
## RMSEA index =  0.117  and the 90 % confidence intervals are  0.108 0.125
## BIC =  231.57
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##                                                 PA1  PA2  PA5  PA3  PA4
## Correlation of (regression) scores with factors   0.97 0.99 0.94 0.94 0.94
## Multiple R square of scores with factors          0.94 0.98 0.88 0.88 0.88
## Minimum correlation of possible factor scores     0.89 0.97 0.76 0.76 0.75
##
##  Coefficients and bootstrapped confidence intervals
##                 low   PA1 upper   low   PA2 upper   low   PA5 upper   low   PA3
## VERBcomp       0.02  0.16  0.32  0.04  0.09  0.13  0.43  0.60  0.81 -0.04  0.01
## compoundVERBs  0.69  0.79  0.86 -0.12 -0.05  0.01 -0.17 -0.08  0.04 -0.03  0.02
## predorder.m   -0.91 -0.76 -0.62 -0.03  0.02  0.07 -0.11  0.02  0.13 -0.05  0.03
## subj           0.65  0.75  0.82  0.06  0.11  0.17 -0.28 -0.16 -0.03 -0.05  0.00
## VERBfrac.m     0.51  0.61  0.72 -0.09 -0.06 -0.03  0.30  0.43  0.59 -0.10 -0.06
## NEGcount.m    -0.15 -0.11 -0.06 -0.08 -0.05 -0.02 -0.01  0.04  0.09  0.84  0.91
```

```
## NEGcount.v     0.11  0.17  0.22  0.04  0.07  0.12 -0.09 -0.03  0.04  0.72  0.80
## NOUNcount.m   -1.00 -0.89 -0.76  0.04  0.07  0.10 -0.20 -0.09  0.00 -0.16 -0.10
## activity       0.29  0.39  0.53 -0.06 -0.03  0.00  0.44  0.65  0.87 -0.03  0.00
## entropy        0.04  0.10  0.14  0.68  0.71  0.75 -0.11 -0.06 -0.01 -0.02  0.01
## hpoint        -0.16 -0.13 -0.10  0.96  0.99  1.01 -0.01  0.04  0.08  0.03  0.06
## maentropy     -0.14 -0.08 -0.02 -0.14 -0.11 -0.08 -0.10 -0.03  0.04 -0.02  0.01
## mamr           0.63  0.74  0.84 -0.08 -0.04  0.01 -0.15 -0.02  0.13 -0.11 -0.05
## hapaxes        0.12  0.18  0.23 -0.91 -0.88 -0.86 -0.07 -0.01  0.05 -0.12 -0.09
## sentcount      0.14  0.22  0.31  0.77  0.80  0.84  0.01  0.09  0.18 -0.19 -0.15
## verbdist      -0.78 -0.69 -0.62 -0.03  0.00  0.03 -0.46 -0.29 -0.16 -0.12 -0.07
##                upper   low   PA4 upper
## VERBcomp       0.06 -0.07 -0.01  0.05
## compoundVERBs  0.08 -0.06  0.00  0.05
## predorder.m    0.12 -0.18 -0.12 -0.05
## subj           0.06 -0.21 -0.14 -0.07
## VERBfrac.m    -0.03 -0.06 -0.03  0.01
## NEGcount.m     1.01 -0.04  0.00  0.04
## NEGcount.v     0.87 -0.02  0.02  0.07
## NOUNcount.m   -0.06 -0.06 -0.02  0.02
## activity       0.04 -0.10 -0.06 -0.02
## entropy        0.05  0.49  0.55  0.59
## hpoint         0.08 -0.07 -0.05 -0.02
## maentropy      0.05  0.72  0.77  0.85
## mamr           0.00 -0.32 -0.26 -0.19
## hapaxes       -0.05  0.24  0.29  0.33
## sentcount     -0.11  0.02  0.05  0.10
## verbdist      -0.01 -0.15 -0.10 -0.06
##
##  Interfactor correlations and bootstrapped confidence intervals
##          lower estimate upper
## PA1-PA2  0.040    0.183  0.33
## PA1-PA5 -0.675    0.604  0.81
## PA1-PA3 -0.656   -0.168  0.90
## PA1-PA4 -0.616   -0.261  0.35
## PA2-PA5  0.028    0.069  0.43
## PA2-PA3 -0.068    0.289  0.36
## PA2-PA4 -0.079    0.162  0.30
## PA5-PA3 -0.462   -0.173  0.36
## PA5-PA4 -0.370   -0.155  0.48
## PA3-PA4 -0.408    0.284  0.41
```

Healthiness diagnostics

```
fa_res$loadings[] %>%
  as_tibble() %>%
  mutate(feat = colnames(data_engineered_2)) %>%
  select(feat, everything()) %>%
  pivot_longer(!feat) %>%
  mutate(value = abs(value)) %>%
  group_by(feat) %>%
  summarize(maxload = max(value)) %>%
  arrange(maxload)
```

```
## # A tibble: 16 x 2
```

```
##    feat          maxload
##    <chr>           <dbl>
##  1 VERBcomp        0.596
##  2 VERBfrac.m      0.607
##  3 activity        0.651
##  4 verbdist        0.694
##  5 entropy         0.711
##  6 mamr            0.737
##  7 subj            0.746
##  8 predorder.m     0.756
##  9 maentropy       0.775
## 10 compoundVERBs   0.785
## 11 NEGcount.v      0.800
## 12 sentcount       0.800
## 13 hapaxes         0.884
## 14 NOUNcount.m     0.888
## 15 NEGcount.m      0.907
## 16 hpoint          0.985
```

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

```
##      VERBcomp   predorder.m compoundVERBs          subj     maentropy
##     0.5177499     0.5183563     0.5349995     0.5414778     0.6418761
##    NEGcount.v          mamr      verbdist     sentcount       hapaxes
##     0.6782765     0.7133565     0.7530430     0.7675342     0.7707273
##    NEGcount.m   NOUNcount.m    VERBfrac.m      activity       entropy
##     0.8304061     0.8350764     0.9038775     0.9052379     0.9451047
##        hpoint
##     0.9601740
```

```
fa_res$communality[fa_res$communality < 0.5] %>% names()
```

```
## character(0)
```

```
fa_res$complexity %>% sort()
```

```
## compoundVERBs    NEGcount.m        hpoint   predorder.m   NOUNcount.m
##      1.029554      1.039301      1.050565      1.057554      1.063257
##     maentropy    NEGcount.v      VERBcomp          subj     sentcount
##      1.063520      1.113116      1.193647      1.212563      1.260463
##          mamr       hapaxes      verbdist      activity    VERBfrac.m
##      1.261246      1.317476      1.397513      1.667989      1.864981
##       entropy
##      1.941602
```

```
fa_res$complexity[fa_res$complexity > 2] %>% names()
```

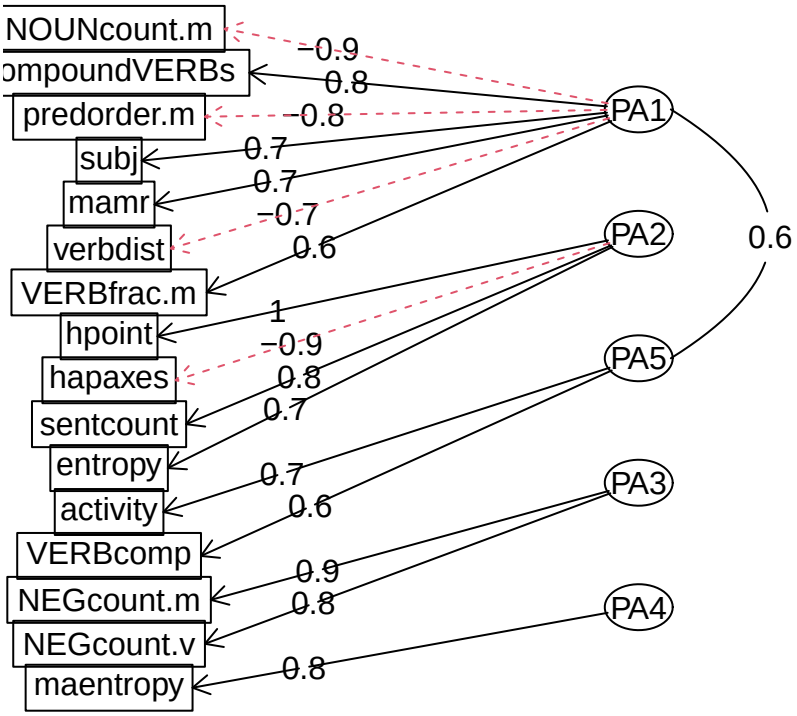
```
## character(0)
```

Loadings

Comrey and Lee (1992): loadings excelent > .70 > very good > .63 > good > .55 > fair > .45 > poor > .32

```
fa.diagram(fa_res)
```

Factor Analysis




```
fa_res$loadings
```

```
##
## Loadings:
##                PA1    PA2    PA5    PA3    PA4
## VERBcomp        0.159          0.596
## compoundVERBs   0.785
## predorder.m    -0.756                        -0.120
## subj            0.746  0.115 -0.157         -0.139
## VERBfrac.m      0.607          0.432
## NEGcount.m     -0.110                 0.907
## NEGcount.v      0.171                 0.800
## NOUNcount.m    -0.888                -0.103
## activity        0.391          0.651
## entropy                0.711                 0.547
## hpoint         -0.134  0.985
## maentropy              -0.109                 0.775
## mamr            0.737                        -0.258
## hapaxes         0.179 -0.884                  0.286
## sentcount       0.217  0.800         -0.150
## verbdist       -0.694         -0.285
##
##                  PA1   PA2   PA5   PA3   PA4
## SS loadings    4.259 2.954 1.103 1.519 1.102
## Proportion Var 0.266 0.185 0.069 0.095 0.069
## Cumulative Var 0.266 0.451 0.520 0.615 0.684
```

```
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 > 0.70 ~ "****",
      abs_l <= 0.70 & abs_l > 0.63 ~ "*** ",
      abs_l <= 0.63 & abs_l > 0.55 ~ "**  ",
      abs_l <= 0.55 & abs_l > 0.45 ~ "*   ",
      abs_l <= 0.45 & abs_l > 0.32 ~ ".   ",
      .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
## NOUNcount.m    -0.888 0.888 ****
## compoundVERBs   0.785 0.785 ****
## predorder.m    -0.756 0.756 ****
## subj            0.746 0.746 ****
## mamr            0.737 0.737 ****
## verbdist       -0.694 0.694 ***
## VERBfrac.m      0.607 0.607 **
## activity        0.391 0.391 .
## sentcount       0.217 0.217
## hapaxes         0.179 0.179
## NEGcount.v      0.171 0.171
## VERBcomp        0.159 0.159
## hpoint         -0.134 0.134
## NEGcount.m     -0.110 0.110
##
##
## ----- PA2 -----
##           loading abs_l  str
## hpoint      0.985 0.985 ****
## hapaxes    -0.884 0.884 ****
## sentcount   0.800 0.800 ****
## entropy     0.711 0.711 ****
## subj        0.115 0.115
## maentropy  -0.109 0.109
##
##
## ----- PA5 -----
```

```
##            loading abs_l  str
## activity     0.651 0.651 ***
## VERBcomp     0.596 0.596 **
## VERBfrac.m   0.432 0.432 .
## verbdist    -0.285 0.285
## subj        -0.157 0.157
##
##
## ----- PA3 -----
##             loading abs_l  str
## NEGcount.m    0.907 0.907 ****
## NEGcount.v    0.800 0.800 ****
## sentcount    -0.150 0.150
## NOUNcount.m  -0.103 0.103
##
##
## ----- PA4 -----
##             loading abs_l  str
## maentropy     0.775 0.775 ****
## entropy       0.547 0.547 *
## hapaxes       0.286 0.286
## mamr         -0.258 0.258
## subj         -0.139 0.139
## predorder.m  -0.120 0.120
```

hypotheses:

- **PA1:** register – narrativity, richness of expression; shorter clauses (-technical / +narrative)
  - long nominal constr., predicate far down, verbs far apart / compound verbs, overt subjects, morphologically diverse, more verbs, activity
- **PA2:** text length (-short / +long)
  - hapaxes load negatively, because I normed them over word count
- **PA5:** activity (-passive / +active)
  - more adjectives / many verbs, more verbcomps
  - nothing to do with compound verbs
  - but something to do with verbal complements
  - `UPOS` of passives annotated as `ADJ` in UD
- **PA3:** negations (-less negated / +more negated)
- **PA4:** lexical richness (-poor / +rich)

strong correlations (but not necessarily significant):

- **PA1+PA5** (-0.67 / **+0.60** / +0.81): narrative texts are active, technical texts are passive

significant correlations (CIs not spanning over 0):

- **PA1+PA2** (+0.10 / **+0.18** / +0.26): narrative texts tend to be slightly longer
  - strange? but the correlation isn't as strong
- **PA2+PA5** (+0.00 / **+0.07** / +0.45): longer texts are more active
  - PA2 behavior opposite to what one would expect

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.

Uniquenesses

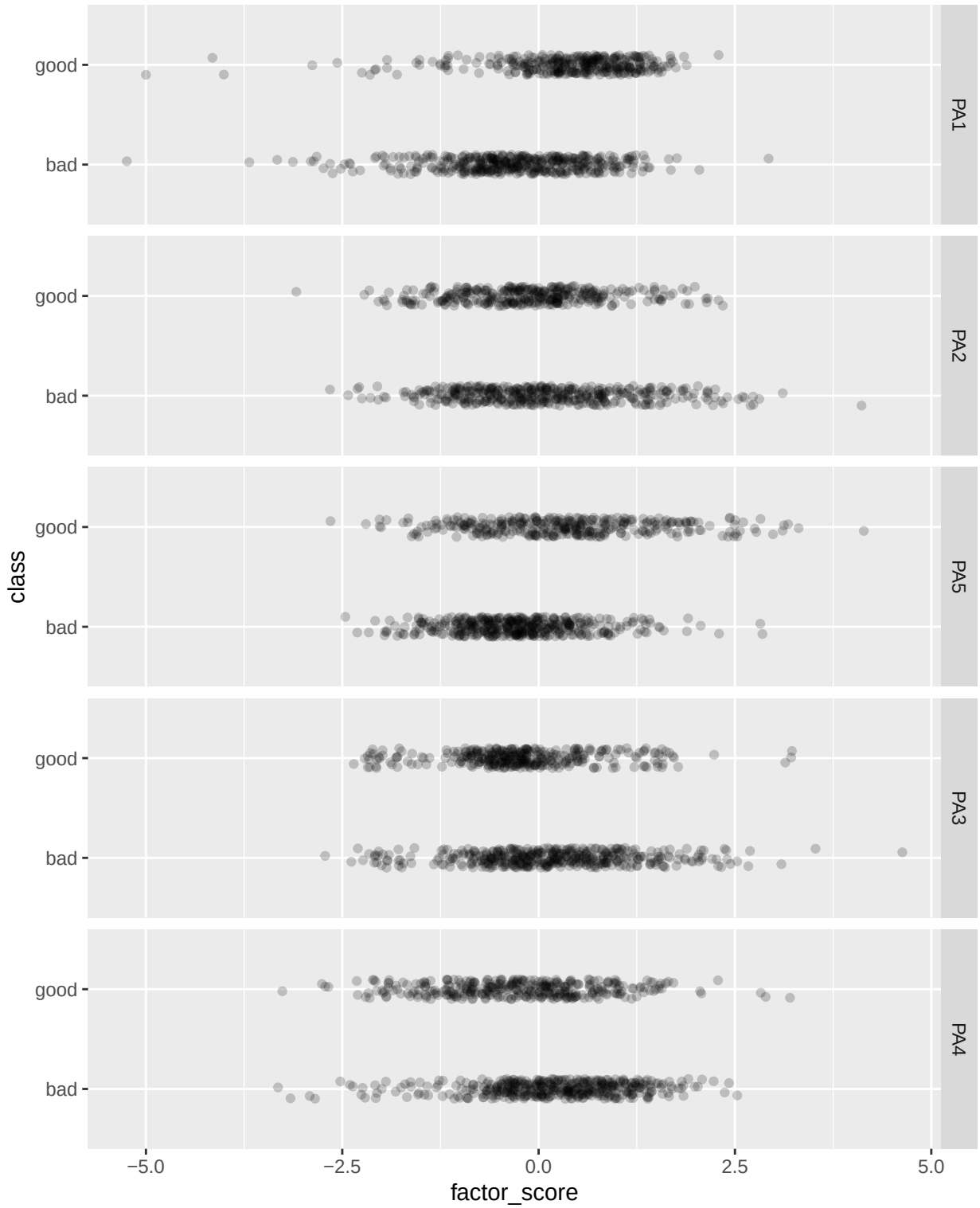
```
fa_res$uniquenesses %>% round(3)
```

```
##      VERBcomp compoundVERBs   predorder.m          subj    VERBfrac.m
##         0.482         0.465         0.482         0.459         0.096
##    NEGcount.m    NEGcount.v   NOUNcount.m      activity       entropy
##         0.170         0.322         0.165         0.095         0.055
##        hpoint     maentropy          mamr       hapaxes     sentcount
##         0.040         0.358         0.287         0.229         0.232
##      verbdist
##         0.247
```

Distributions over factors

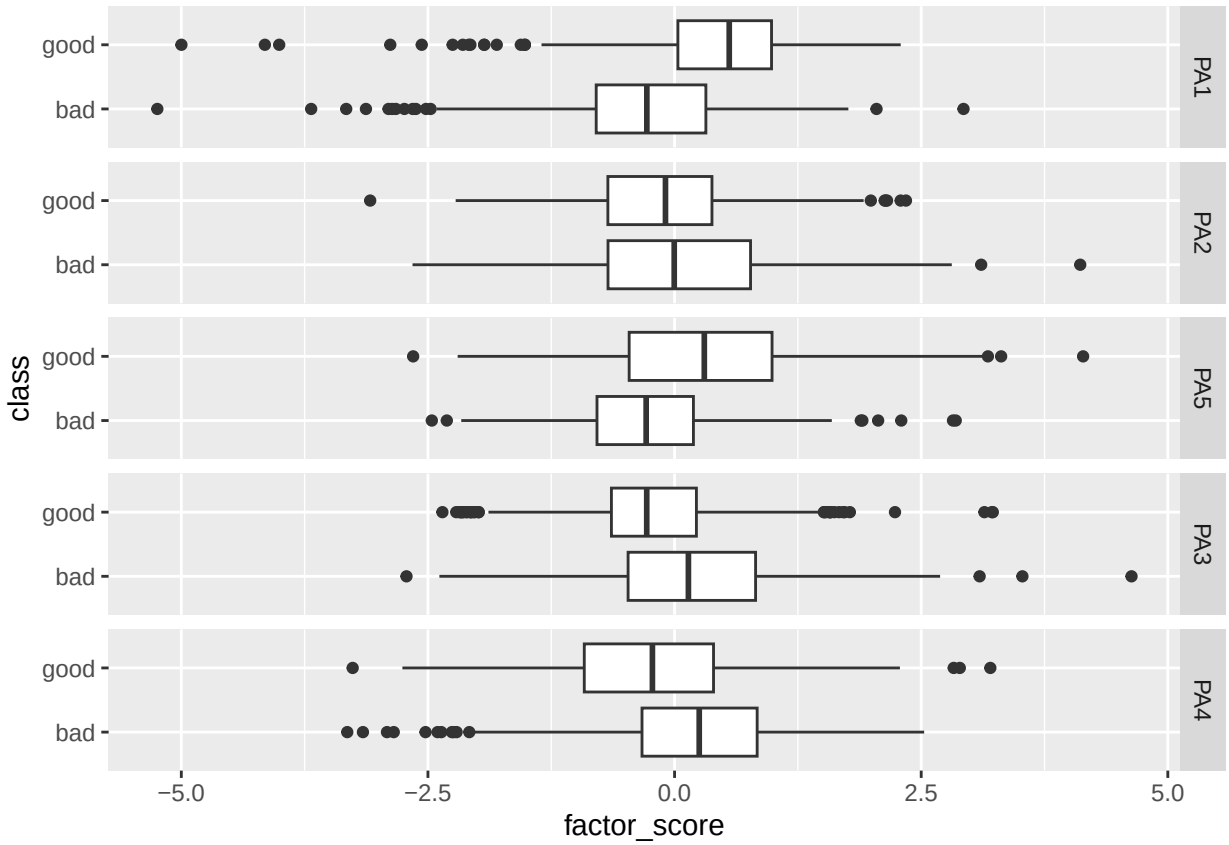
```
analyze_distributions <- function(data_factors_long, variable) {
  plot <- data_factors_long %>%
    ggplot(aes(x = factor_score, y = !!sym(variable))) +
    geom_boxplot() +
    facet_grid(factor ~ .)
  print(plot)

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

  p_val <- numeric()
  epsilon2 <- numeric()
  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"
    )

    kw <- kruskal.test(data$factor_score, data[[variable]])

    dunn <- dunn.test(
      data$factor_score, data[[variable]],
      altp = TRUE, method = "bonferroni"
    )

    e2 <- epsilonSquared(data$factor_score, data[[variable]])
    cat("epsilon2 = ", e2, "\n")

    min_p_values <- c(min_p_values, min(dunn$altP.adjusted))
    p_val <- c(p_val, kw$p.value)
    epsilon2 <- c(epsilon2, e2)
  }

  cat("\n")
  print(data.frame(factor = factors, kruskal_p = p_val, epsilon2 = epsilon2), digits = 3)
```

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

data_factors <- bind_cols(data_clean, fa_res$scores %>% as.data.frame())
cnames <- map(
  colnames(data_factors),
  function(x) {
    name <- pull(pretty_names %>%
      filter(name_orig == x), name_pretty)
    if (length(name) == 1) {
      return(name)
    } else {
      return(x)
    }
  }
) %>% unlist()
colnames(data_factors) <- cnames

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", "PA5", "PA3", "PA4"))
  ))

data_factors_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_long, "class")
```



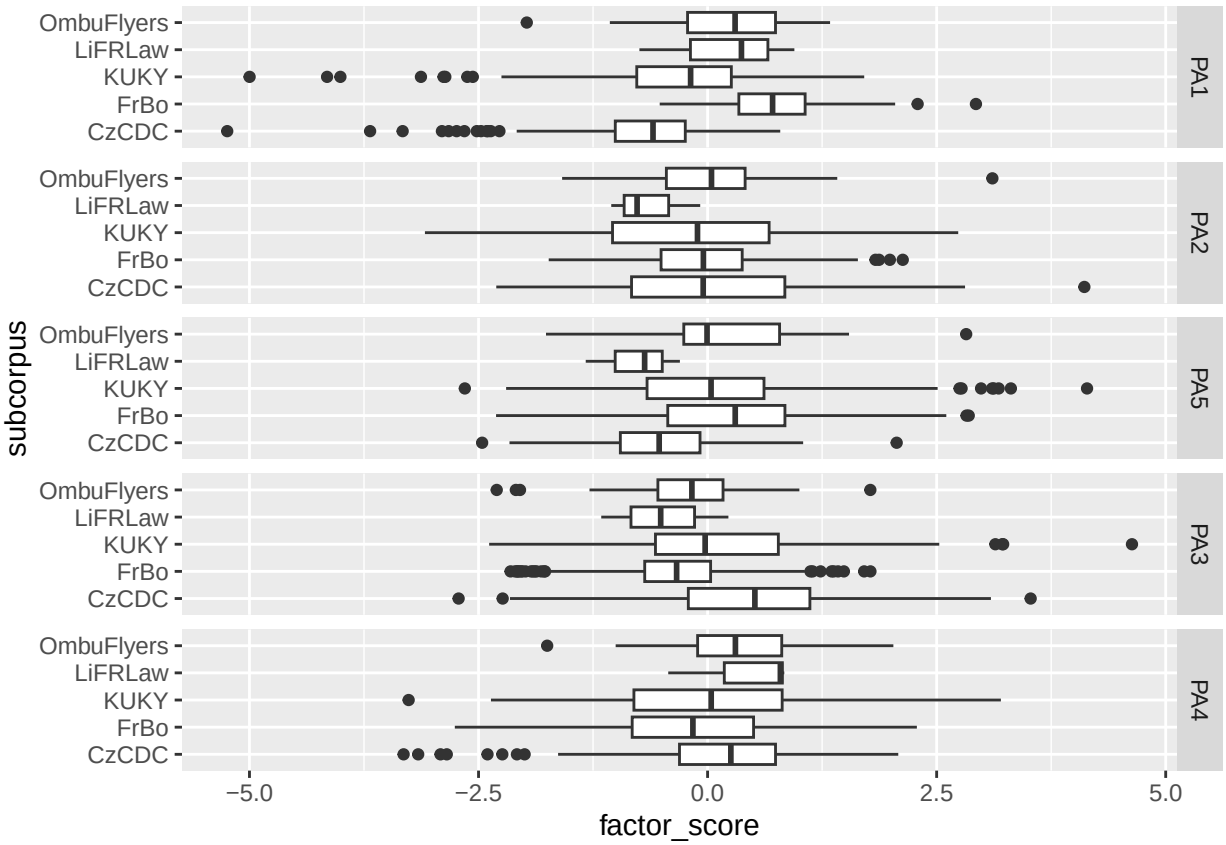

```
##
## Test for the significance of differences in class over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 126.7269, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |        bad
## ---------+-----------
##     good |  -11.25730
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.168
##
## Test for the significance of differences in class over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 4.3681, df = 1, p-value = 0.04
```

```
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |        bad
## ---------+-----------
##     good |   2.089988
##          |    0.0366*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0058
##
## Test for the significance of differences in class over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 66.1797, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |        bad
## ---------+-----------
##     good |  -8.135091
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0879
##
## Test for the significance of differences in class over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 32.512, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |        bad
## ---------+-----------
##     good |   5.701925
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0432
##
```

```
## Test for the significance of differences in class over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 42.0912, df = 1, p-value = 0
##
##
##                             Comparison of x by group
##                                   (Bonferroni)
## Col Mean-|
## Row Mean |        bad
## ---------+-----------
##     good |   6.487771
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0559
##
##   factor kruskal_p epsilon2
## 1    PA1  2.13e-29   0.1680
## 2    PA2  3.66e-02   0.0058
## 3    PA5  4.12e-16   0.0879
## 4    PA3  1.18e-08   0.0432
## 5    PA4  8.71e-11   0.0559
##
## p < 5e-2 found in: PA1 PA2 PA5 PA3 PA4
## p < 1e-2 found in: PA1 PA5 PA3 PA4
## p < 1e-3 found in: PA1 PA5 PA3 PA4
## p < 1e-4 found in: PA1 PA5 PA3 PA4
```

subcorpus

```
analyze_distributions(data_factors_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 = 367.4879, df = 4, p-value = 0
##
##
##                             Comparison of x by group
##                                   (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY    LiFRLaw
## ---------+--------------------------------------------
##     FrBo |  -18.12656
##          |    0.0000*
##          |
##     KUKY |  -4.397196   12.79333
##          |    0.0001*    0.0000*
##          |
##  LiFRLaw |  -1.688090   1.090683  -0.935521
##          |     0.9139     1.0000     1.0000
##          |
## OmbuFlye |  -5.873484   3.374214  -3.361550  -0.087571
##          |    0.0000*    0.0074*    0.0078*     1.0000
##
## alpha = 0.05
```

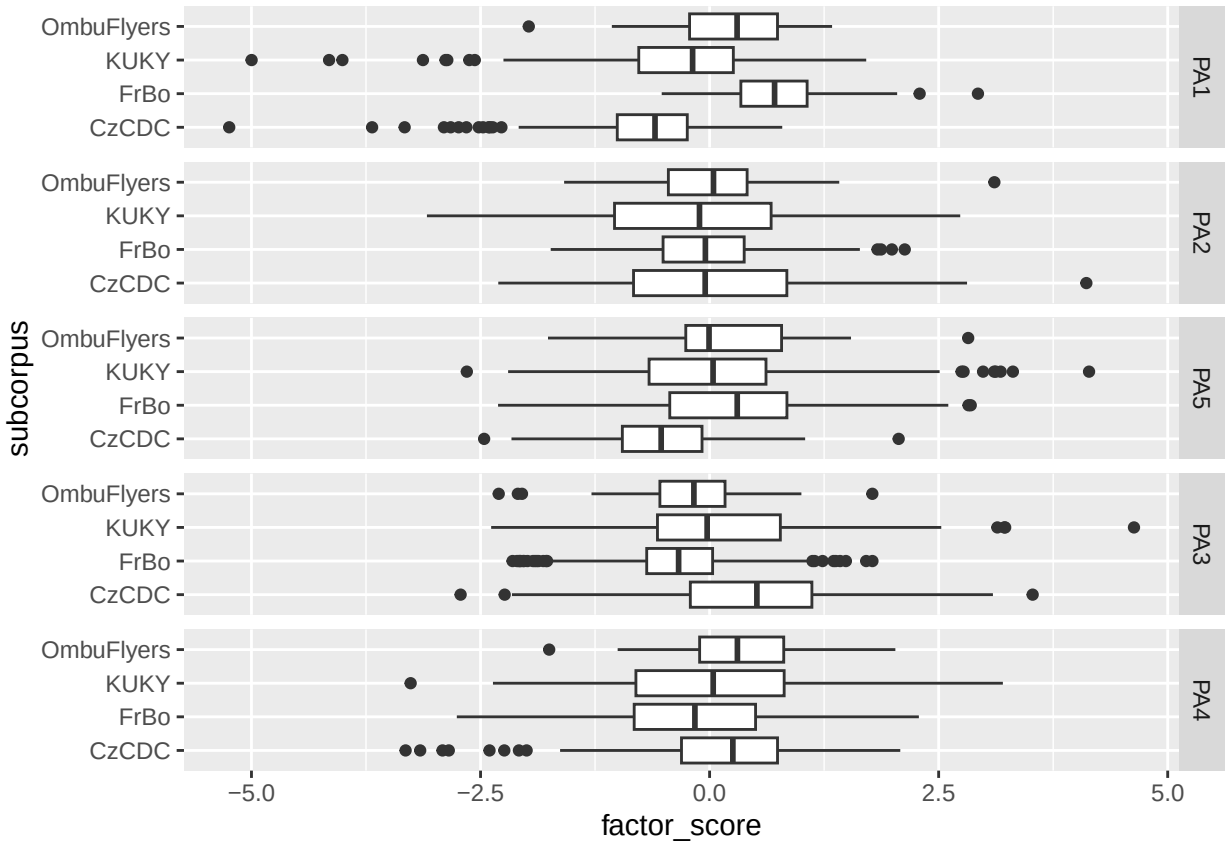
```
## Reject Ho if p <= alpha
## epsilon2 =  0.488
##
## Test for the significance of differences in subcorpus over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 4.895, df = 4, p-value = 0.3
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY    LiFRLaw
## ---------+-----------------------------------------------
##     FrBo |   0.691349
##          |     1.0000
##          |
##     KUKY |   1.649817   1.113294
##          |     0.9898     1.0000
##          |
##  LiFRLaw |   1.400227   1.297050   1.117258
##          |     1.0000     1.0000     1.0000
##          |
## OmbuFlye |  -0.234238  -0.597786  -1.155916  -1.426199
##          |     1.0000     1.0000     1.0000     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0065
##
## Test for the significance of differences in subcorpus over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 110.959, df = 4, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY    LiFRLaw
## ---------+-----------------------------------------------
##     FrBo |  -10.16261
##          |    0.0000*
##          |
##     KUKY |  -6.687235   2.611257
##          |    0.0000*     0.0902
##          |
##  LiFRLaw |   0.571656   2.132759   1.713677
##          |     1.0000     0.3294     0.8659
##          |
```

```
## OmbuFlye |  -4.799832   0.349228  -1.014940  -1.963114
##          |    0.0000*     1.0000     1.0000     0.4963
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.147
##
## Test for the significance of differences in subcorpus over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 100.0432, df = 4, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY    LiFRLaw
## ---------+--------------------------------------------
##     FrBo |   9.887361
##          |    0.0000*
##          |
##     KUKY |   4.675599  -4.518170
##          |    0.0000*    0.0001*
##          |
##  LiFRLaw |   1.855136   0.341381   1.054870
##          |     0.6358     1.0000     1.0000
##          |
## OmbuFlye |   3.769452  -1.261441   1.119595  -0.691979
##          |    0.0016*     1.0000     1.0000     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.133
##
## Test for the significance of differences in subcorpus over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 24.5286, df = 4, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY    LiFRLaw
## ---------+--------------------------------------------
##     FrBo |   4.228071
##          |    0.0002*
##          |
##     KUKY |   2.074281  -1.851174
##          |     0.3805     0.6414
```

```
##          |
##  LiFRLaw |  -0.423701  -1.073575  -0.777744
##          |     1.0000     1.0000     1.0000
##          |
## OmbuFlye |  -1.086128  -3.301276  -2.238183   0.091937
##          |     1.0000    0.0096*     0.2521     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0326
##
##   factor kruskal_p epsilon2
## 1    PA1  2.93e-78   0.4880
## 2    PA2  2.98e-01   0.0065
## 3    PA5  4.54e-23   0.1470
## 4    PA3  9.63e-21   0.1330
## 5    PA4  6.26e-05   0.0326
##
## p < 5e-2 found in: PA1 PA5 PA3 PA4
## p < 1e-2 found in: PA1 PA5 PA3 PA4
## p < 1e-3 found in: PA1 PA5 PA3 PA4
## p < 1e-4 found in: PA1 PA5 PA3
```

subcorpus wo/ LiFRLaw

```
analyze_distributions(
  data_factors_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 = 367.2784, df = 3, p-value = 0
##
##
##                             Comparison of x by group
##                                   (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY
## ---------+---------------------------------
##     FrBo |  -18.12400
##          |    0.0000*
##          |
##     KUKY |  -4.398355   12.78960
##          |    0.0001*    0.0000*
##          |
## OmbuFlye |  -5.870726   3.375713  -3.358166
##          |    0.0000*    0.0044*    0.0047*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.49
##
```

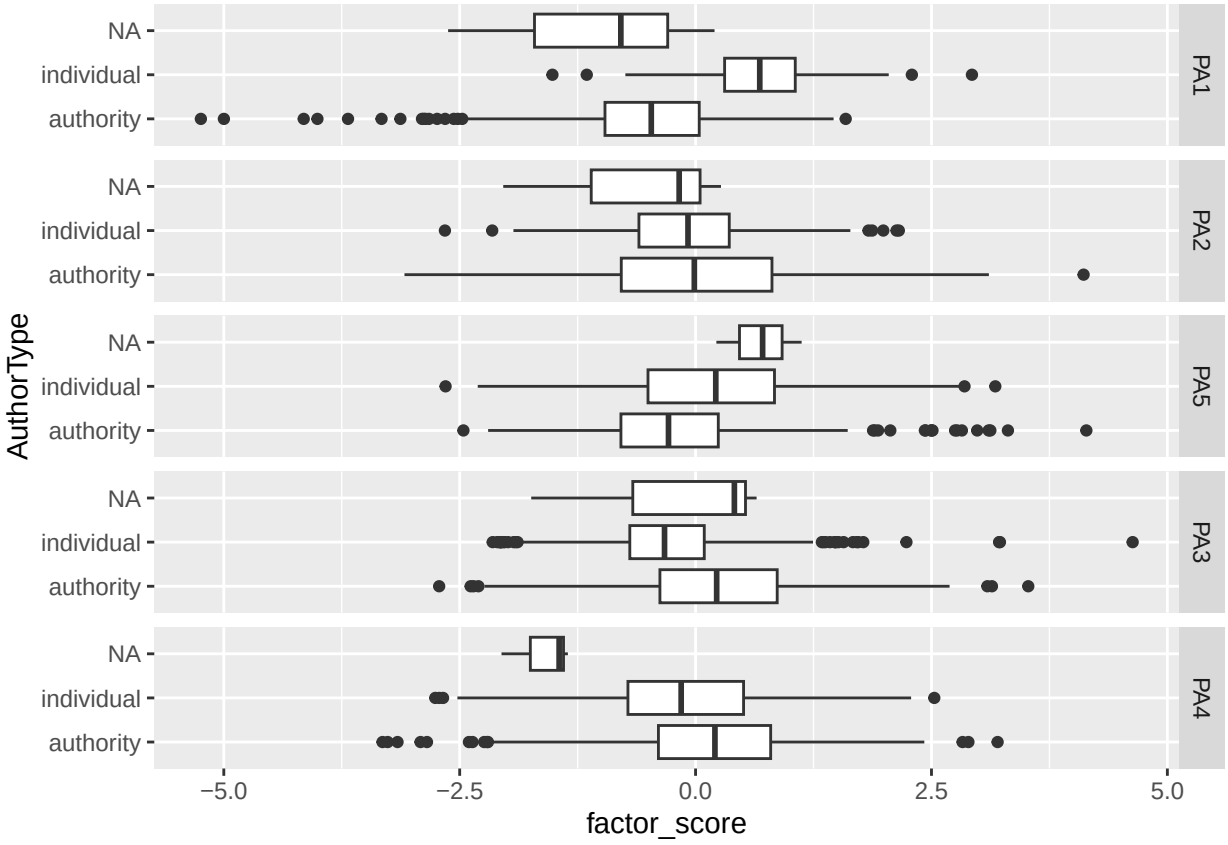
```
## Test for the significance of differences in subcorpus over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 3.2066, df = 3, p-value = 0.36
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY
## ---------+---------------------------------
##     FrBo |   0.707561
##          |     1.0000
##          |
##     KUKY |   1.650663   1.098516
##          |     0.5928     1.0000
##          |
## OmbuFlye |  -0.225616  -0.597355  -1.147841
##          |     1.0000     1.0000     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.00428
##
## Test for the significance of differences in subcorpus over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 108.5966, df = 3, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY
## ---------+---------------------------------
##     FrBo |  -10.16932
##          |    0.0000*
##          |
##     KUKY |  -6.695230   2.609106
##          |    0.0000*     0.0545
##          |
## OmbuFlye |  -4.798707   0.353850  -1.009348
##          |    0.0000*     1.0000     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.145
##
## Test for the significance of differences in subcorpus over PA3 :
##
```

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 99.1033, df = 3, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY
## ---------+---------------------------------
##     FrBo |   9.888077
##          |    0.0000*
##          |
##     KUKY |   4.673653  -4.520964
##          |    0.0000*    0.0000*
##          |
## OmbuFlye |   3.769553  -1.261709   1.120783
##          |    0.0010*     1.0000     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.132
##
## Test for the significance of differences in subcorpus over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 23.972, df = 3, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY
## ---------+---------------------------------
##     FrBo |   4.230854
##          |    0.0001*
##          |
##     KUKY |   2.078297  -1.849528
##          |     0.2261     0.3863
##          |
## OmbuFlye |  -1.086490  -3.303088  -2.240791
##          |     1.0000    0.0057*     0.1502
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.032
##
##   factor kruskal_p epsilon2
## 1    PA1  2.70e-79  0.49000
## 2    PA2  3.61e-01  0.00428
## 3    PA5  2.20e-23  0.14500
```

```
## 4    PA3  2.42e-21  0.13200
## 5    PA4  2.53e-05  0.03200
##
## p < 5e-2 found in: PA1 PA5 PA3 PA4
## p < 1e-2 found in: PA1 PA5 PA3 PA4
## p < 1e-3 found in: PA1 PA5 PA3 PA4
## p < 1e-4 found in: PA1 PA5 PA3
```

AuthorType

```
analyze_distributions(data_factors_long, "AuthorType")
```



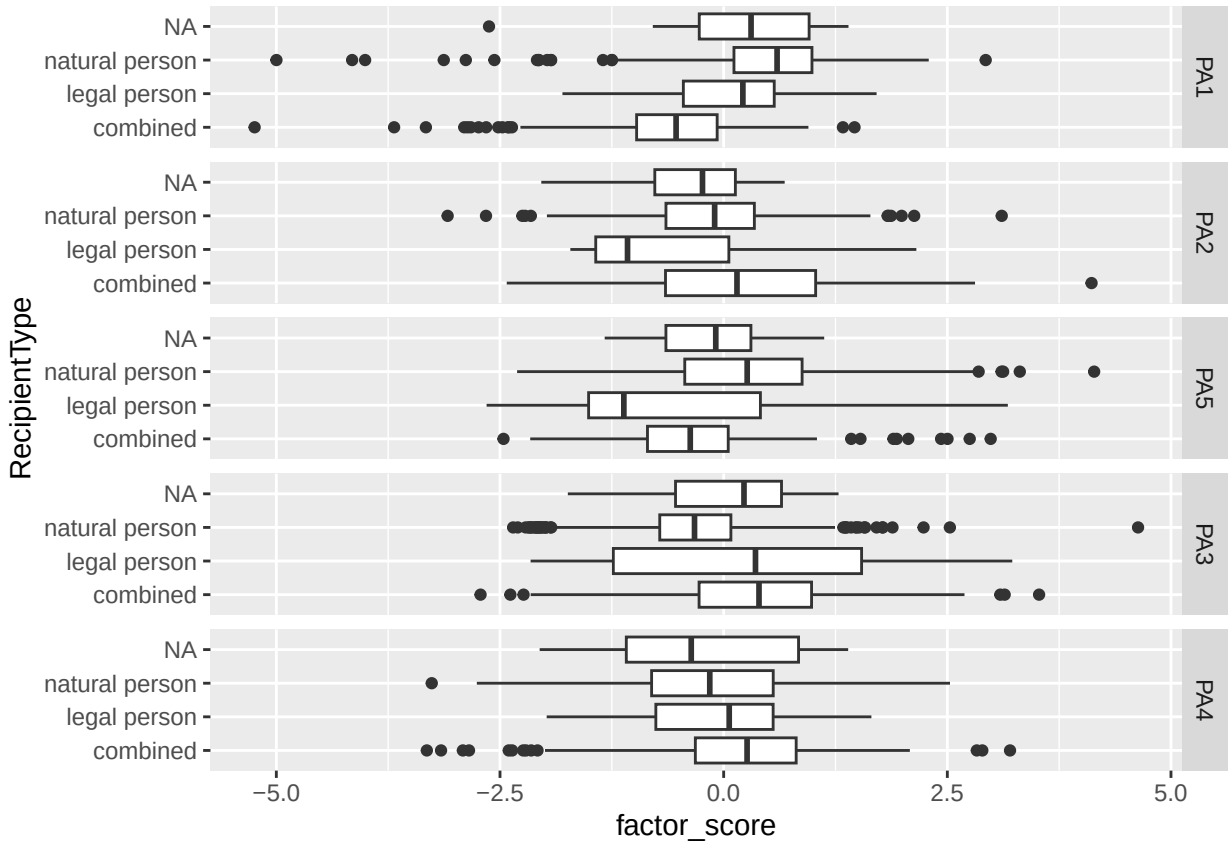

```
##
## Test for the significance of differences in AuthorType over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 340.431, df = 1, p-value = 0
##
##
##                              Comparison of x by group
##                                    (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## ---------+-----------
## individu |  -18.45077
```

```
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.452
##
## Test for the significance of differences in AuthorType over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 1.7006, df = 1, p-value = 0.19
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## ---------+-----------
## individu |   1.304053
##          |     0.1922
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.00226
##
## Test for the significance of differences in AuthorType over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 43.8958, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## ---------+-----------
## individu |  -6.625390
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0583
##
## Test for the significance of differences in AuthorType over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 60.8503, df = 1, p-value = 0
##
##
```

```
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## ---------+-----------
## individu |   7.800661
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0808
##
## Test for the significance of differences in AuthorType over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 17.5505, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## ---------+-----------
## individu |   4.189331
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0233
##
##   factor kruskal_p epsilon2
## 1    PA1  5.14e-76  0.45200
## 2    PA2  1.92e-01  0.00226
## 3    PA5  3.46e-11  0.05830
## 4    PA3  6.16e-15  0.08080
## 5    PA4  2.80e-05  0.02330
##
## p < 5e-2 found in: PA1 PA5 PA3 PA4
## p < 1e-2 found in: PA1 PA5 PA3 PA4
## p < 1e-3 found in: PA1 PA5 PA3 PA4
## p < 1e-4 found in: PA1 PA5 PA3 PA4
```

RecipientType

```
analyze_distributions(data_factors_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 = 274.7923, df = 2, p-value = 0
##
##
##                             Comparison of x by group
##                                   (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## ---------+----------------------
## legal pe |  -3.556014
##          |    0.0011*
##          |
## natural  |  -16.57496  -2.247529
##          |    0.0000*     0.0738
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.365
##
## Test for the significance of differences in RecipientType over PA2 :
##
##   Kruskal-Wallis rank sum test
```

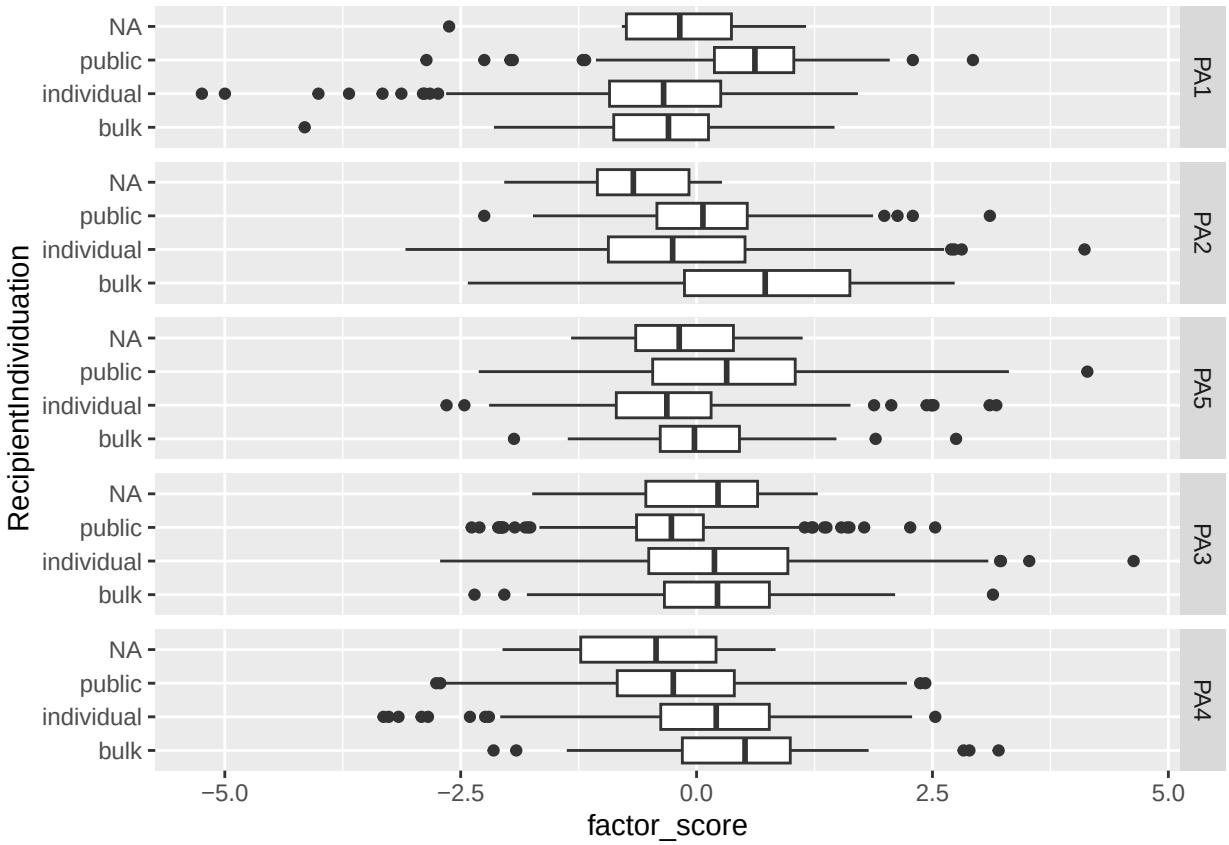
```
##
## data: x and group
## Kruskal-Wallis chi-squared = 23.3807, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## ---------+----------------------
## legal pe |   3.882907
##          |    0.0003*
##          |
## natural  |   3.599127  -2.650511
##          |    0.0010*    0.0241*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0311
##
## Test for the significance of differences in RecipientType over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 92.961, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## ---------+----------------------
## legal pe |   0.192234
##          |     1.0000
##          |
## natural  |  -9.407062  -3.505768
##          |    0.0000*    0.0014*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.123
##
## Test for the significance of differences in RecipientType over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 101.3913, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
```

```
## Row Mean |   combined   legal pe
## ---------+----------------------
## legal pe |   1.259825
##          |     0.6232
##          |
## natural  |   10.04052   2.263746
##          |    0.0000*     0.0708
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.135
##
## Test for the significance of differences in RecipientType over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 21.9911, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## ---------+----------------------
## legal pe |   1.256302
##          |     0.6270
##          |
## natural  |   4.677943   0.379369
##          |    0.0000*     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0292
##
##   factor kruskal_p epsilon2
## 1    PA1  2.14e-60   0.3650
## 2    PA2  8.37e-06   0.0311
## 3    PA5  6.51e-21   0.1230
## 4    PA3  9.62e-23   0.1350
## 5    PA4  1.68e-05   0.0292
##
## p < 5e-2 found in: PA1 PA2 PA5 PA3 PA4
## p < 1e-2 found in: PA1 PA2 PA5 PA3 PA4
## p < 1e-3 found in: PA1 PA2 PA5 PA3 PA4
## p < 1e-4 found in: PA1 PA5 PA3 PA4
```

court decisions often with `RecipientType = combined`.

RecipientIndividuation

```
analyze_distributions(data_factors_long, "RecipientIndividuation")
```



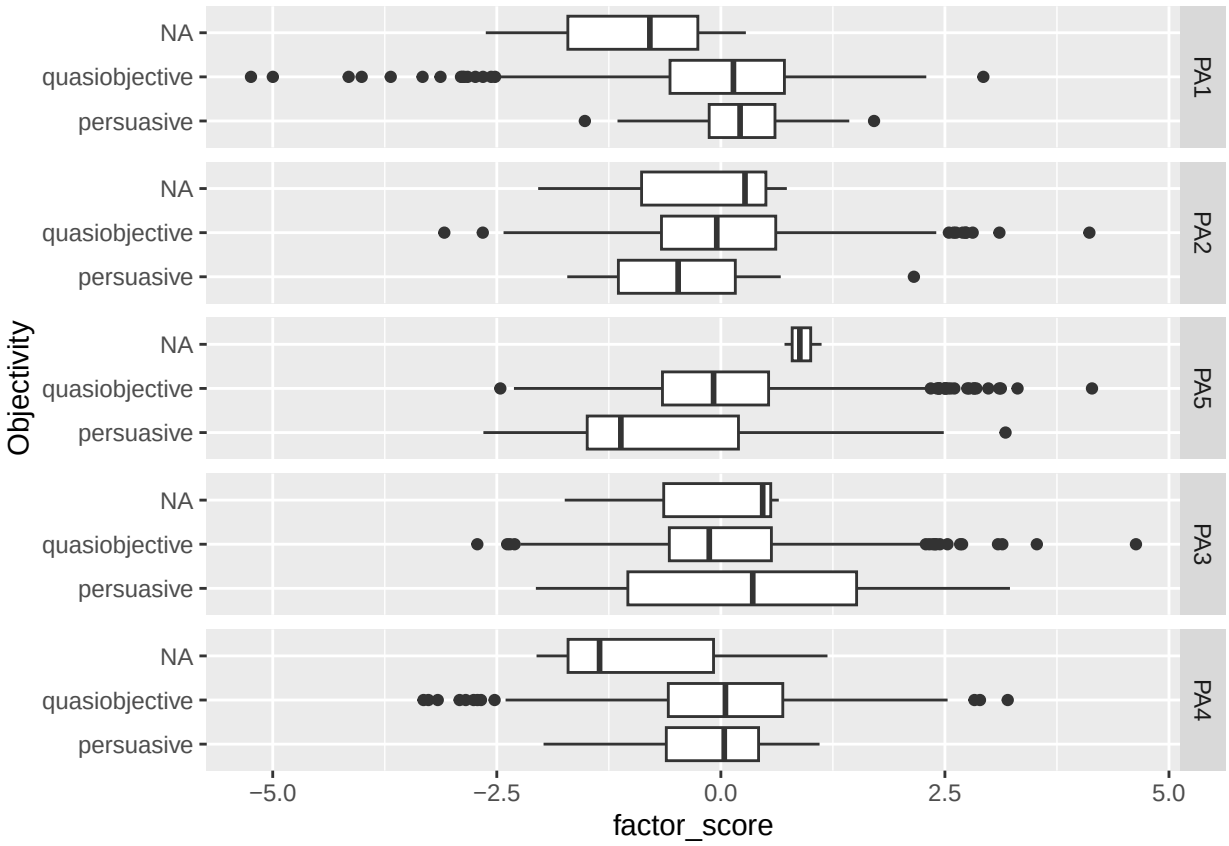

```
##
## Test for the significance of differences in RecipientIndividuation over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 213.054, df = 2, p-value = 0
##
##
##                             Comparison of x by group
##                                   (Bonferroni)
## Col Mean-|
## Row Mean |       bulk   individu
## ---------+----------------------
## individu |  -0.658353
##          |     1.0000
##          |
##   public |  -8.680142  -13.83849
##          |    0.0000*    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.283
##
## Test for the significance of differences in RecipientIndividuation over PA2 :
##
##   Kruskal-Wallis rank sum test
```

```
##
## data: x and group
## Kruskal-Wallis chi-squared = 39.4788, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |       bulk   individu
## ---------+----------------------
## individu |   5.840692
##          |    0.0000*
##          |
##   public |   3.555540  -3.832922
##          |    0.0011*    0.0004*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0524
##
## Test for the significance of differences in RecipientIndividuation over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 72.9615, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |       bulk   individu
## ---------+----------------------
## individu |   2.905428
##          |    0.0110*
##          |
##   public |  -2.069169  -8.521702
##          |     0.1156    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0969
##
## Test for the significance of differences in RecipientIndividuation over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 45.7366, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
```

```
## Row Mean |       bulk   individu
## ---------+----------------------
## individu |   0.536624
##          |     1.0000
##          |
##   public |   4.205896   6.334174
##          |    0.0001*    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0607
##
## Test for the significance of differences in RecipientIndividuation over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 39.7899, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |       bulk   individu
## ---------+----------------------
## individu |   1.758924
##          |     0.2358
##          |
##   public |   4.838917   5.340624
##          |    0.0000*    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0528
##
##   factor kruskal_p epsilon2
## 1    PA1  5.44e-47   0.2830
## 2    PA2  2.67e-09   0.0524
## 3    PA5  1.43e-16   0.0969
## 4    PA3  1.17e-10   0.0607
## 5    PA4  2.29e-09   0.0528
##
## p < 5e-2 found in: PA1 PA2 PA5 PA3 PA4
## p < 1e-2 found in: PA1 PA2 PA5 PA3 PA4
## p < 1e-3 found in: PA1 PA2 PA5 PA3 PA4
## p < 1e-4 found in: PA1 PA2 PA5 PA3 PA4
```

Objectivity

```
analyze_distributions(data_factors_long, "Objectivity")
```



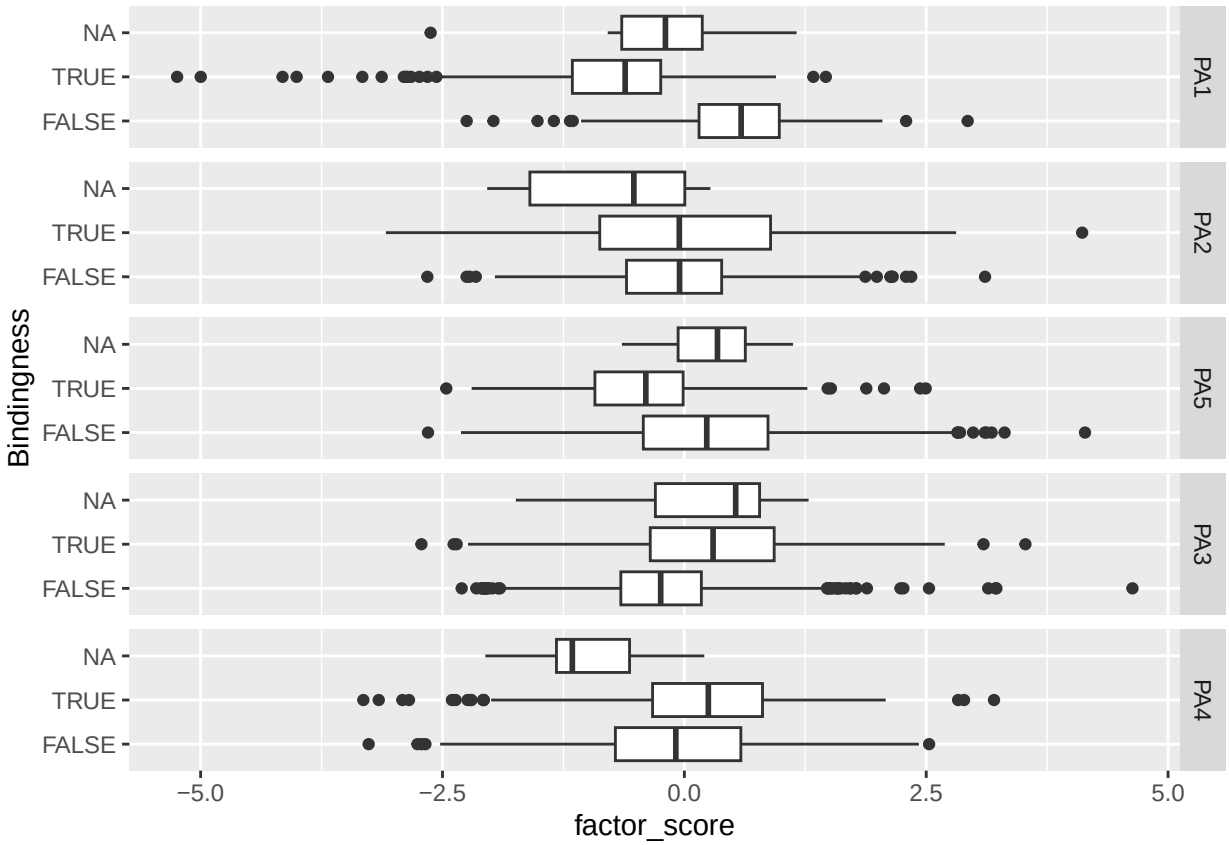

```
##
## Test for the significance of differences in Objectivity over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.5086, df = 1, p-value = 0.48
##
##
##                             Comparison of x by group
##                                   (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## ---------+-----------
## quasiobj |   0.713161
##          |     0.4757
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.000675
##
## Test for the significance of differences in Objectivity over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 5.3827, df = 1, p-value = 0.02
```

```
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## ---------+-----------
## quasiobj |  -2.320070
##          |    0.0203*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.00715
##
## Test for the significance of differences in Objectivity over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 5.8222, df = 1, p-value = 0.02
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## ---------+-----------
## quasiobj |  -2.412913
##          |    0.0158*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.00773
##
## Test for the significance of differences in Objectivity over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.5808, df = 1, p-value = 0.45
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## ---------+-----------
## quasiobj |   0.762133
##          |     0.4460
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.000771
##
```

```
## Test for the significance of differences in Objectivity over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.3873, df = 1, p-value = 0.53
##
##
##                             Comparison of x by group
##                                   (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## ---------+-----------
## quasiobj |  -0.622358
##          |     0.5337
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.000514
##
##   factor kruskal_p epsilon2
## 1    PA1    0.4757 0.000675
## 2    PA2    0.0203 0.007150
## 3    PA5    0.0158 0.007730
## 4    PA3    0.4460 0.000771
## 5    PA4    0.5337 0.000514
##
## p < 5e-2 found in: PA2 PA5
## p < 1e-2 found in:
## p < 1e-3 found in:
## p < 1e-4 found in:
```

Bindingness

```
analyze_distributions(data_factors_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 = 356.1166, df = 1, p-value = 0
##
##
##                             Comparison of x by group
##                                   (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## ---------+-----------
##     TRUE |   18.87105
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.473
##
## Test for the significance of differences in Bindingness over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.8725, df = 1, p-value = 0.35
```

```
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## ---------+-----------
##     TRUE |  -0.934087
##          |     0.3503
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.00116
##
## Test for the significance of differences in Bindingness over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 97.2718, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## ---------+-----------
##     TRUE |   9.862645
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.129
##
## Test for the significance of differences in Bindingness over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 52.7326, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## ---------+-----------
##     TRUE |  -7.261724
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.07
##
```

```
## Test for the significance of differences in Bindingness over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 17.2115, df = 1, p-value = 0
##
##
##                             Comparison of x by group
##                                   (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## ---------+-----------
##     TRUE |  -4.148671
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0229
##
##   factor kruskal_p epsilon2
## 1    PA1  1.97e-79  0.47300
## 2    PA2  3.50e-01  0.00116
## 3    PA5  6.04e-23  0.12900
## 4    PA3  3.82e-13  0.07000
## 5    PA4  3.34e-05  0.02290
##
## p < 5e-2 found in: PA1 PA5 PA3 PA4
## p < 1e-2 found in: PA1 PA5 PA3 PA4
## p < 1e-3 found in: PA1 PA5 PA3 PA4
## p < 1e-4 found in: PA1 PA5 PA3 PA4
```

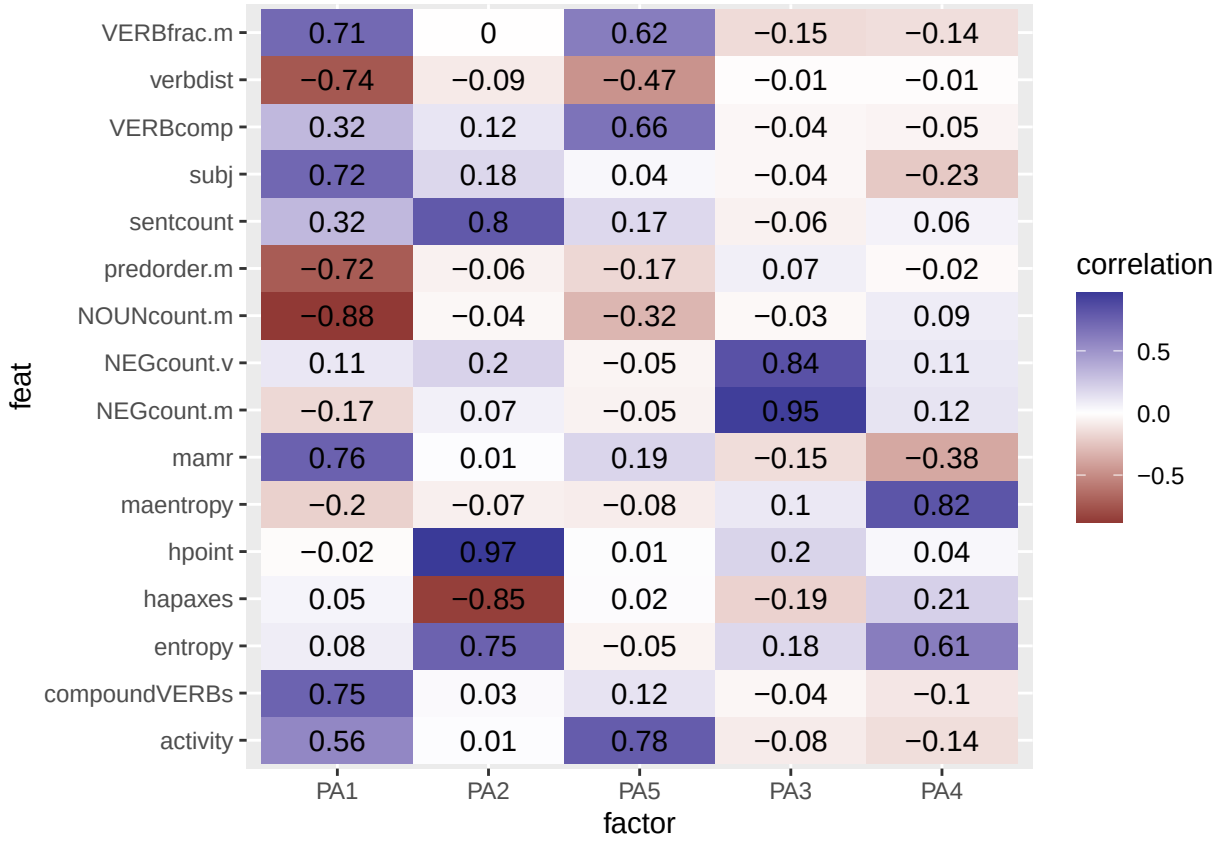
Feature-factor correlations

```
data_factors_longer <- data_factors_long %>%
  pivot_longer(
    abstractNOUNs:verbdist,
    names_to = "feat", values_to = "feat_value"
  )

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

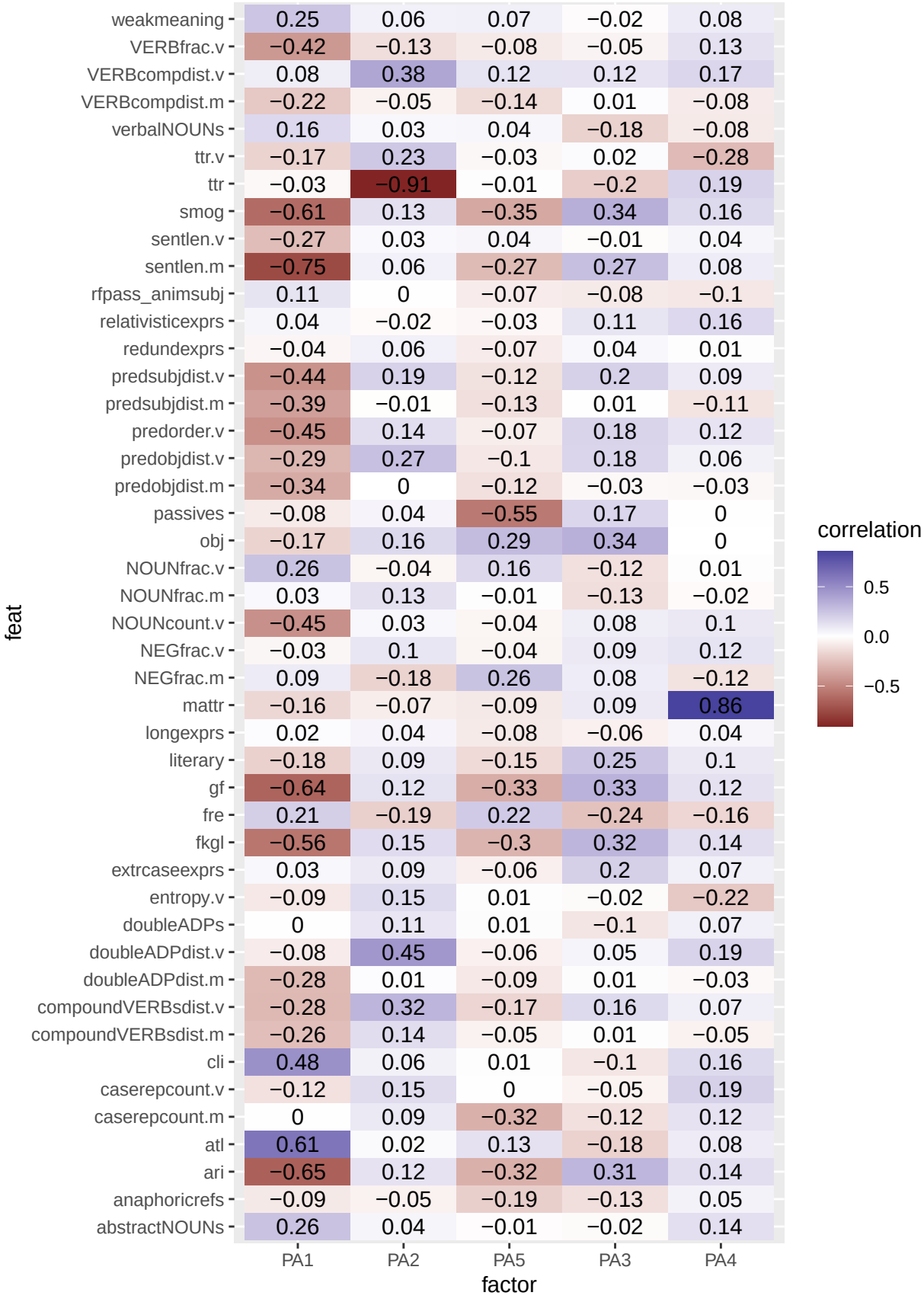
```
## `summarise()` has grouped output by 'feat'. You can override using the
## `.groups` argument.
```

```
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()
```




```
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
weakmeaning 0.25 0.06 0.07 −0.02 0.08
VERBfrac.v −0.42 −0.13 −0.08 −0.05 0.13
VERBcompdist.v 0.08 0.38 0.12 0.12 0.17
VERBcompdist.m −0.22 −0.05 −0.14 0.01 −0.08
verbalNOUNs 0.16 0.03 0.04 −0.18 −0.08
ttr.v −0.17 0.23 −0.03 0.02 −0.28
ttr −0.03 −0.91 −0.01 −0.2 0.19
smog −0.61 0.13 −0.35 0.34 0.16
sentlen.v −0.27 0.03 0.04 −0.01 0.04
sentlen.m −0.75 0.06 −0.27 0.27 0.08
rfpass_animsubj 0.11 0 −0.07 −0.08 −0.1
relativisticexprs 0.04 −0.02 −0.03 0.11 0.16
redundexprs −0.04 0.06 −0.07 0.04 0.01
predsubjdist.v −0.44 0.19 −0.12 0.2 0.09
predsubjdist.m −0.39 −0.01 −0.13 0.01 −0.11
predorder.v −0.45 0.14 −0.07 0.18 0.12
predobjdist.v −0.29 0.27 −0.1 0.18 0.06
predobjdist.m −0.34 0 −0.12 −0.03 −0.03
passives −0.08 0.04 −0.55 0.17 0
obj −0.17 0.16 0.29 0.34 0
NOUNfrac.v 0.26 −0.04 0.16 −0.12 0.01
NOUNfrac.m 0.03 0.13 −0.01 −0.13 −0.02
NOUNcount.v −0.45 0.03 −0.04 0.08 0.1
NEGfrac.v −0.03 0.1 −0.04 0.09 0.12
NEGfrac.m 0.09 −0.18 0.26 0.08 −0.12
mattr −0.16 −0.07 −0.09 0.09 0.86
longexprs 0.02 0.04 −0.08 −0.06 0.04
literary −0.18 0.09 −0.15 0.25 0.1
gf −0.64 0.12 −0.33 0.33 0.12
fre 0.21 −0.19 0.22 −0.24 −0.16
fkgl −0.56 0.15 −0.3 0.32 0.14
extrcaseexprs 0.03 0.09 −0.06 0.2 0.07
entropy.v −0.09 0.15 0.01 −0.02 −0.22
doubleADPs 0 0.11 0.01 −0.1 0.07
doubleADPdist.v −0.08 0.45 −0.06 0.05 0.19
doubleADPdist.m −0.28 0.01 −0.09 0.01 −0.03
compoundVERBsdist.v −0.28 0.32 −0.17 0.16 0.07
compoundVERBsdist.m −0.26 0.14 −0.05 0.01 −0.05
cli 0.48 0.06 0.01 −0.1 0.16
caserepcount.v −0.12 0.15 0 −0.05 0.19
caserepcount.m 0 0.09 −0.32 −0.12 0.12
atl 0.61 0.02 0.13 −0.18 0.08
ari −0.65 0.12 −0.32 0.31 0.14
anaphoricrefs −0.09 −0.05 −0.19 −0.13 0.05
abstractNOUNs 0.26 0.04 −0.01 −0.02 0.14
PA1 PA2 PA5 PA3 PA4
factor
correlation
0.5
0.0
−0.5