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
```

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

```
## Loading required package: lattice

##
## Attaching package: 'lattice'

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

##
## Attaching package: 'nFactors'

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

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

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

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

## The following object is masked from 'package:rcompanion':
##
##     phi
```

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

```
## Loading required package: ggplot2

##
## Attaching package: 'ggplot2'

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

##
## Attaching package: 'caret'

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

```r
library(tidyverse)
```

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

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

```
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: 753 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
```

```
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,
```

```r
  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
))
```

```r
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 text-length dependent
  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 further variables belonging to the 'acceptability' category
  select(!c(RuleIncompleteConjunction)) %>%
  # remove artificially limited variables
  select(!c(
    RuleCaseRepetition.max_repetition_frac,
    RuleCaseRepetition.max_repetition_frac.v
  )) %>%
  # remove variables with too many NAs
  select(!c(
    RuleDoubleAdpos.max_allowable_distance,
    RuleDoubleAdpos.max_allowable_distance.v
  )) %>%
  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 77
## # i 77 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: 61 x 2
##    feat_name                                  p_value
##    <chr>                                        <dbl>
##  1 RuleAbstractNouns                          2.20e- 3
##  2 RuleAnaphoricReferences                    6.73e- 1
##  3 RuleCaseRepetition.max_repetition_count    6.59e- 2
##  4 RuleCaseRepetition.max_repetition_count.v 4.54e- 3
##  5 RuleConfirmationExpressions                1.08e- 1
##  6 RuleDoubleAdpos                            2.71e- 1
##  7 RuleInfVerbDistance                        5.24e-15
##  8 RuleInfVerbDistance.max_distance           5.48e- 2
##  9 RuleInfVerbDistance.max_distance.v         6.58e- 2
## 10 RuleLiteraryStyle                          7.00e-21
## # i 51 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                          0.943   0.943
##  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 0.943   0.943
##  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                         0.948   0.948
## 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                   0.958   0.958
## 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                         0.948   0.948
## 22 word_count                        hpoint                       0.958   0.958
```

exclude:

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

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

```
## Compare row 5  and column  32 with corr  0.943
##   Means:  0.407 vs 0.214 so flagging column 5
## Compare row 32  and column  38 with corr  0.978
##   Means:  0.388 vs 0.206 so flagging column 32
## Compare row 38  and column  46 with corr  0.987
##   Means:  0.374 vs 0.199 so flagging column 38
## Compare row 46  and column  36 with corr  0.948
##   Means:  0.353 vs 0.191 so flagging column 46
## Compare row 33  and column  34 with corr  0.96
##   Means:  0.265 vs 0.187 so flagging column 33
## Compare row 40  and column  43 with corr  0.964
##   Means:  0.179 vs 0.184 so flagging column 43
## Compare row 48  and column  39 with corr  0.958
##   Means:  0.185 vs 0.184 so flagging column 48
## All correlations <= 0.9
```

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

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

```r
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

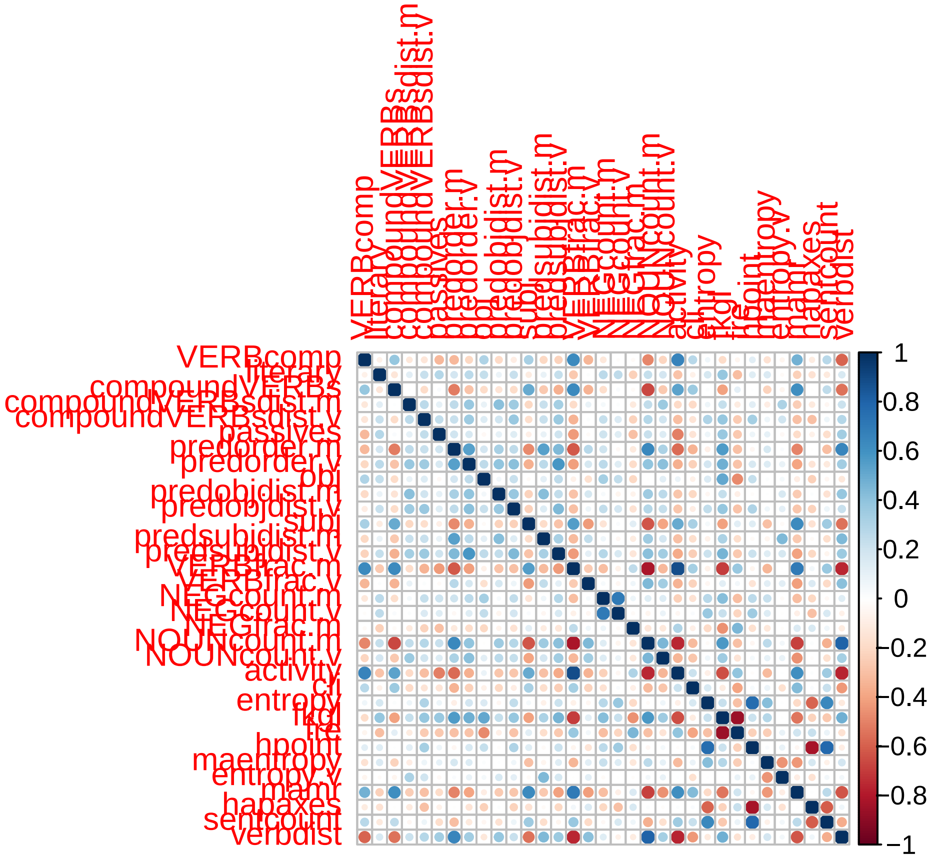
```
# 0.35 instead of 0.3 otherwise the FA bootstrapping would freeze
.lcorrcutoff <- 0.35

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                               0.00220
## 2 RuleCaseRepetition.max_repetition_count.v       0.00454
## 3 RuleRedundantExpressions                        0.0103
## 4 RuleRelativisticExpressions                     0.00199
## 5 RuleTooManyNegations.max_negation_frac.v        0.0323
## 6 RuleTooManyNominalConstructions.max_noun_frac.v 0.00000482
## 7 RuleVerbalNouns                                 0.000115
## 8 RuleWeakMeaningWords                            0.0490
## 9 GPs                                             0.0144
```

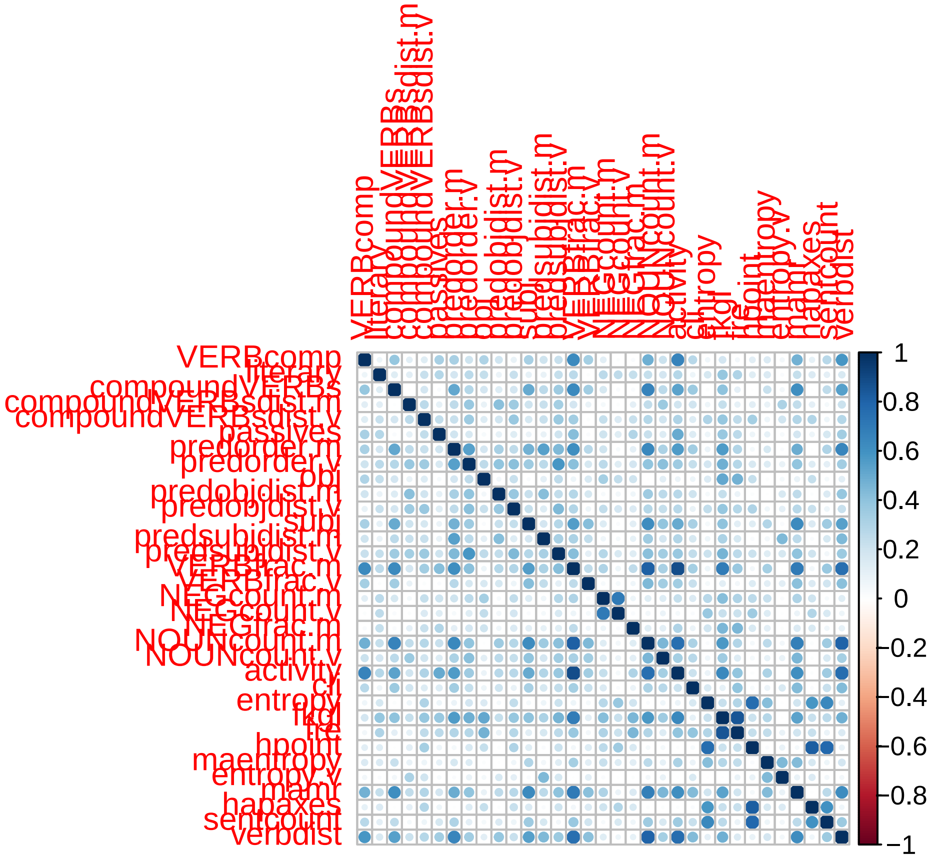
```
data_pure <- data_pureish_striphigh %>%
  select(!any_of(low_correlating_features))

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

colnames(data_pure) <- cnames
```

Visualisation

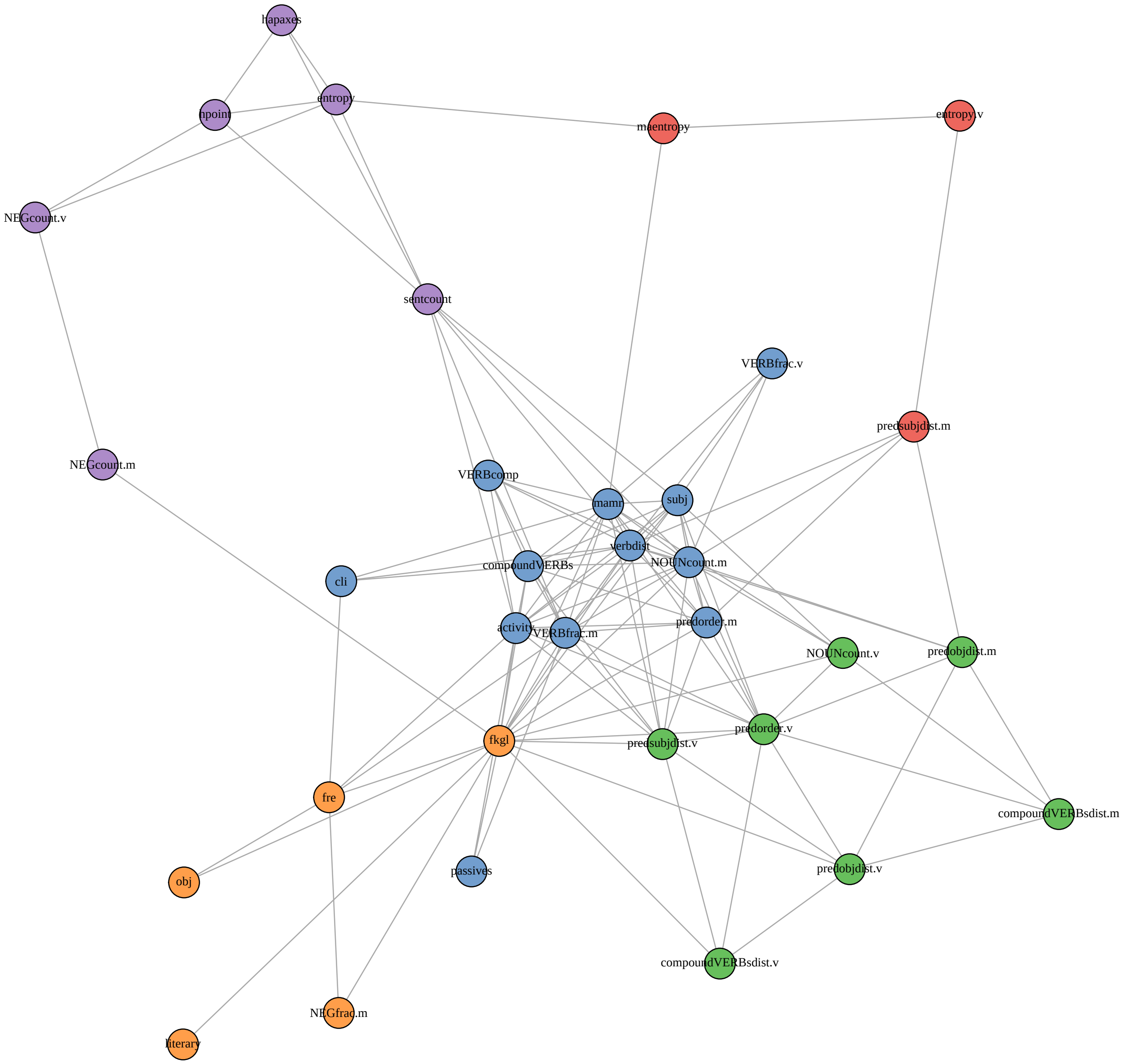
```
corrplot(cor(data_pure))
```



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



```r
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
```

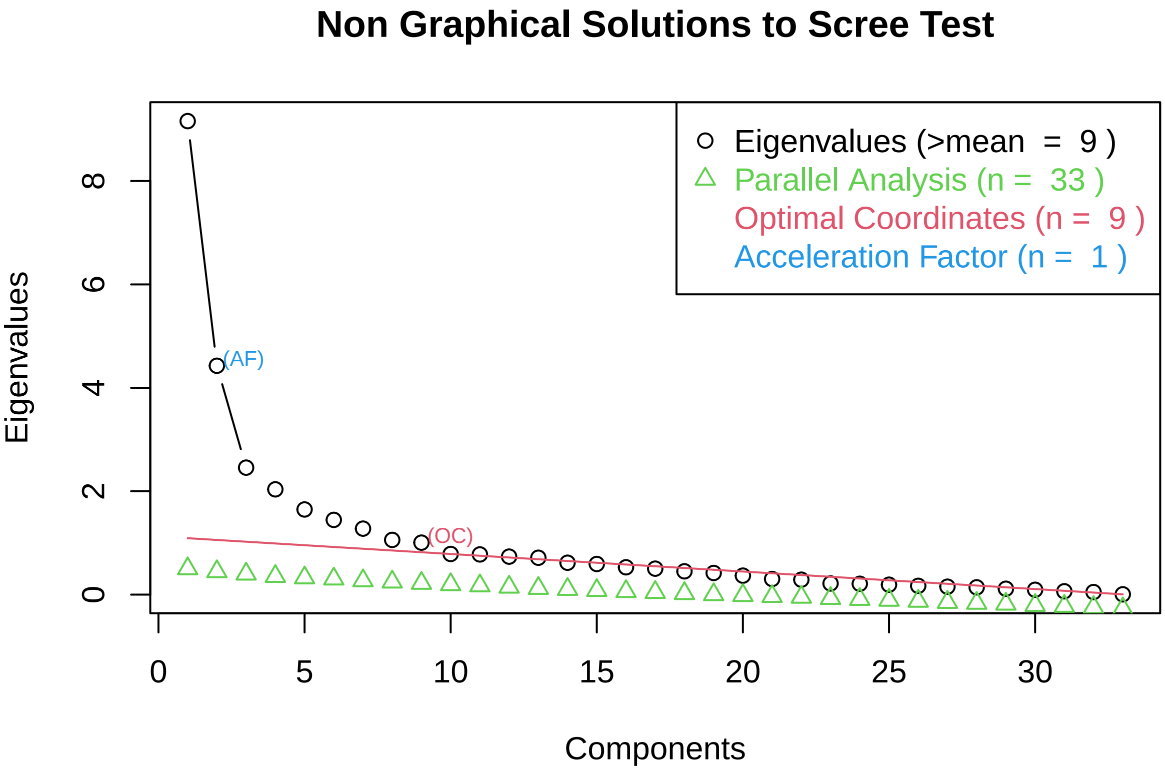
```
##          Beta-hat       kappa p-val
## Skewness 1072.176 134558.0274     0
## Kurtosis 2721.144    447.0881     0
```

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

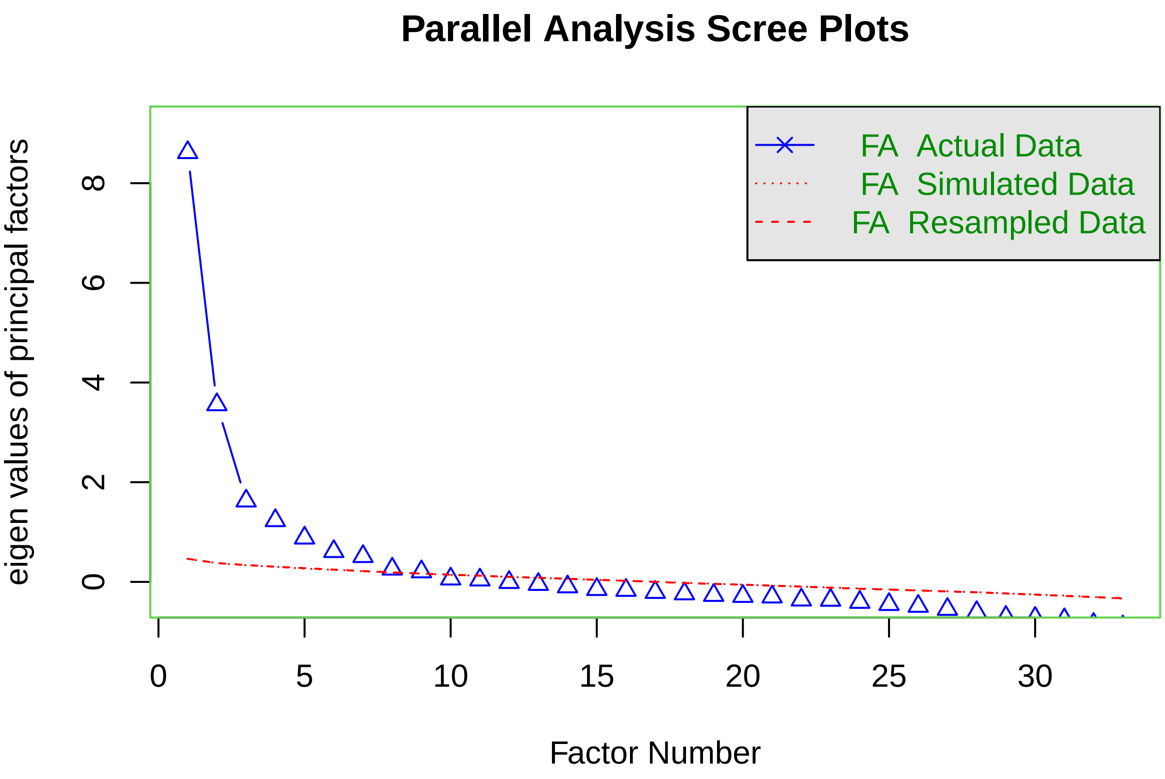
No. of factors

```
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

```
set.seed(42)

# 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
)
```

```
## maximum iteration exceeded

## 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.
```

```
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   PA8   PA3   PA9   h2
## VERBcomp             0.60  0.02  0.05  0.50  0.29 -0.12  0.06 -0.03  0.00 0.59
## literary             0.01 -0.04  0.08  0.16 -0.29  0.14 -0.03 -0.03  0.07 0.24
## compoundVERBs        1.04 -0.14  0.32 -0.28 -0.31  0.04  0.16  0.00 -0.01 0.72
## compoundVERBsdist.m  0.23 -0.04  0.74 -0.08 -0.10 -0.07 -0.05  0.13  0.08 0.46
## compoundVERBsdist.v -0.10  0.24  0.25  0.01 -0.19  0.03 -0.03  0.08  0.02 0.33
## passives             0.06 -0.08  0.02 -0.22 -0.82  0.09 -0.09 -0.04 -0.06 0.55
## predorder.m         -0.56 -0.05  0.32  0.17  0.13 -0.05  0.00 -0.16 -0.26 0.69
## predorder.v         -0.03  0.00  0.72  0.12  0.05  0.04 -0.04 -0.15  0.03 0.58
## obj                  0.13 -0.06  0.00  0.91  0.15  0.12 -0.08  0.04 -0.06 0.70
## predobjdist.m        0.02 -0.09  0.71 -0.13  0.05 -0.06  0.06 -0.02 -0.10 0.41
## predobjdist.v        0.05  0.15  0.56  0.04 -0.01  0.05  0.03 -0.05  0.03 0.38
## subj                 0.64  0.14 -0.11 -0.07 -0.09  0.06  0.08  0.02 -0.29 0.56
## predsubjdist.m      -0.30 -0.05  0.25  0.05  0.14  0.03  0.18  0.31 -0.29 0.48
## predsubjdist.v      -0.17  0.11  0.47  0.13  0.02  0.08 -0.06 -0.02 -0.01 0.46
## VERBfrac.m           0.85 -0.05  0.14  0.03  0.35 -0.01  0.06  0.05  0.02 0.90
## VERBfrac.v          -0.56 -0.07  0.05 -0.18  0.22  0.02  0.04  0.14  0.20 0.37
## NEGcount.m           0.00 -0.10 -0.04  0.17  0.05  0.99  0.00  0.02 -0.04 0.97
## NEGcount.v           0.20  0.07 -0.02  0.06 -0.05  0.73  0.04  0.06  0.07 0.58
## NEGfrac.m           -0.03 -0.03  0.01 -0.21  0.50  0.28 -0.11 -0.06 -0.15 0.42
## NOUNcount.m         -0.89  0.03  0.06 -0.03 -0.02 -0.14  0.02 -0.03  0.04 0.81
## NOUNcount.v         -0.20 -0.08  0.39  0.05 -0.02 -0.02 -0.11  0.03  0.26 0.37
## activity             0.76 -0.01  0.10  0.25  0.49  0.01 -0.10  0.01 -0.01 0.93
## cli                  0.37  0.00 -0.06 -0.12  0.10  0.01  0.79 -0.01 -0.01 0.72
## entropy             -0.07  0.75  0.06 -0.13  0.04  0.10  0.19 -0.05  0.39 0.86
## fkgl                -0.39 -0.05 -0.02  0.53 -0.29  0.05  0.18  0.02 -0.01 0.96
## fre                  0.06  0.05  0.05 -0.46  0.20 -0.06 -0.63 -0.02  0.07 0.98
## hpoint              -0.05  0.98  0.02  0.02  0.00 -0.01 -0.06  0.00  0.03 0.95
## maentropy           -0.36  0.02  0.00 -0.10  0.07  0.09  0.23 -0.42  0.43 0.60
## entropy.v           -0.07  0.05 -0.03  0.03  0.05  0.10  0.03  0.97 -0.04 0.92
## mamr                 0.82 -0.04  0.03 -0.05  0.00 -0.04  0.15 -0.06 -0.40 0.80
## hapaxes              0.04 -0.82  0.07 -0.12  0.08  0.02  0.13 -0.09  0.13 0.74
## sentcount            0.15  0.94  0.05 -0.24  0.28 -0.08  0.07 -0.04  0.03 0.88
## verbdist            -0.79  0.00  0.10 -0.22 -0.17 -0.07 -0.07 -0.05 -0.13 0.81
##                        u2 com
## VERBcomp            0.407 2.6
## literary            0.764 2.5
## compoundVERBs       0.278 1.6
## compoundVERBsdist.m 0.540 1.4
## compoundVERBsdist.v 0.674 3.5
## passives            0.446 1.2
## predorder.m         0.306 2.7
## predorder.v         0.424 1.2
## obj                 0.296 1.2
## predobjdist.m       0.592 1.2
## predobjdist.v       0.621 1.2
## subj                0.437 1.7
## predsubjdist.m      0.523 5.1
## predsubjdist.v      0.535 1.7
## VERBfrac.m          0.099 1.4
## VERBfrac.v          0.633 2.1
## NEGcount.m          0.029 1.1
## NEGcount.v          0.415 1.2
## NEGfrac.m           0.577 2.4
## NOUNcount.m         0.187 1.1
## NOUNcount.v         0.632 2.7
## activity            0.075 2.1
## cli                 0.280 1.5
## entropy             0.143 1.8
## fkgl                0.039 2.8
## fre                 0.020 2.2
## hpoint              0.046 1.0
## maentropy           0.398 3.8
## entropy.v           0.085 1.1
## mamr                0.195 1.6
## hapaxes             0.264 1.2
## sentcount           0.117 1.4
## verbdist            0.191 1.4
##
##                       PA1  PA2  PA7  PA4  PA6  PA5  PA8  PA3  PA9
## SS loadings           6.6 3.14 2.55 2.06 2.04 1.71 1.35 1.33 0.96
## Proportion Var        0.2 0.10 0.08 0.06 0.06 0.05 0.04 0.04 0.03
## Cumulative Var        0.2 0.30 0.37 0.43 0.50 0.55 0.59 0.63 0.66
## Proportion Explained  0.3 0.14 0.12 0.09 0.09 0.08 0.06 0.06 0.04
## Cumulative Proportion 0.3 0.45 0.57 0.66 0.75 0.83 0.89 0.96 1.00
##
##  With factor correlations of
##       PA1   PA2   PA7   PA4   PA6   PA5   PA8   PA3   PA9
## PA1  1.00  0.11 -0.61 -0.23  0.41 -0.26 -0.08 -0.02  0.04
## PA2  0.11  1.00  0.16  0.31 -0.24  0.31  0.18  0.04  0.10
## PA7 -0.61  0.16  1.00  0.39 -0.42  0.26  0.06  0.27 -0.01
## PA4 -0.23  0.31  0.39  1.00 -0.42  0.26  0.30 -0.05  0.11
## PA6  0.41 -0.24 -0.42 -0.42  1.00 -0.30 -0.32  0.06 -0.02
## PA5 -0.26  0.31  0.26  0.26 -0.30  1.00  0.03 -0.15  0.11
## PA8 -0.08  0.18  0.06  0.30 -0.32  0.03  1.00 -0.18  0.10
## PA3 -0.02  0.04  0.27 -0.05  0.06 -0.15 -0.18  1.00 -0.12
## PA9  0.04  0.10 -0.01  0.11 -0.02  0.11  0.10 -0.12  1.00
##
## Mean item complexity =  1.9
## Test of the hypothesis that 9 factors are sufficient.
##
## df null model =  528  with the objective function =  27.86 with Chi Square =  20623.86
## df of  the model are 267  and the objective function was  3.86
##
## 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  753 with the empirical chi square  436.04  with prob <  2.8e-10
## The total n.obs was  753  with Likelihood Chi Square =  2833.68  with prob <  0
##
## Tucker Lewis Index of factoring reliability =  0.745
## RMSEA index =  0.113  and the 90 % confidence intervals are  0.109 0.117
## BIC =  1065.05
## Fit based upon off diagonal values = 0.99
##  Coefficients and bootstrapped confidence intervals
##                       low   PA1 upper   low   PA2 upper   low   PA7 upper   low
## VERBcomp            -0.71  0.60  2.02 -0.06  0.02  0.10 -0.21  0.05  0.33 -2.15
## literary            -0.10  0.01  0.10 -0.20 -0.04  0.12 -1.30  0.08  1.63 -0.67
## compoundVERBs       -1.43  1.04  3.65 -1.86 -0.14  1.43 -2.90  0.32  3.85 -1.78
## compoundVERBsdist.m -0.17  0.23  0.60 -0.12 -0.04  0.06 -7.07  0.74  9.45 -0.24
## compoundVERBsdist.v -0.39 -0.10  0.17 -2.18  0.24  2.94 -3.69  0.25  4.69 -0.14
## passives            -0.12  0.06  0.23 -1.04 -0.08  0.79 -0.49  0.02  0.63 -1.47
## predorder.m         -2.20 -0.56  0.92 -1.41 -0.05  1.17 -1.71  0.32  2.48 -0.66
## predorder.v         -0.21 -0.03  0.11 -0.28  0.00  0.26 -7.29  0.72  9.56 -0.40
## obj                 -0.27  0.13  0.59 -0.41 -0.06  0.26 -0.49  0.00  0.60 -3.54
## predobjdist.m       -0.14  0.02  0.16 -1.82 -0.09  1.45 -6.49  0.71  8.72 -0.41
## predobjdist.v       -0.12  0.05  0.17 -0.76  0.15  1.15 -5.00  0.56  6.69 -0.31
## subj                -0.83  0.64  2.24 -0.99  0.14  1.39 -2.41 -0.11  1.92 -0.16
## predsubjdist.m      -0.98 -0.30  0.33 -0.63 -0.05  0.47 -2.66  0.25  3.52 -0.09
## predsubjdist.v      -0.63 -0.17  0.24 -1.06  0.11  1.42 -4.80  0.47  6.32 -0.46
## VERBfrac.m          -1.15  0.85  2.98 -0.37 -0.05  0.25 -1.96  0.14  2.43 -0.04
## VERBfrac.v          -1.86 -0.56  0.62 -0.51 -0.07  0.34 -1.00  0.05  1.20 -1.10
## NEGcount.m          -0.06  0.00  0.05 -1.05 -0.10  0.78 -1.09 -0.04  0.90 -0.78
## NEGcount.v          -0.30  0.20  0.74 -0.64  0.07  0.85 -0.45 -0.02  0.45 -0.08
## NEGfrac.m           -0.23 -0.03  0.14 -0.57 -0.03  0.43 -0.30  0.01  0.19 -1.08
## NOUNcount.m         -3.17 -0.89  1.20 -0.02  0.03  0.07 -0.31  0.06  0.44 -0.16
## NOUNcount.v         -0.42 -0.20 -0.01 -0.70 -0.08  0.50 -6.10  0.39  7.61 -0.07
## activity            -1.08  0.76  2.75 -0.05 -0.01  0.02 -1.68  0.10  2.06 -0.86
## cli                 -0.50  0.37  1.33 -0.06  0.00  0.05 -0.86 -0.06  0.61 -0.86
## entropy             -0.24 -0.07  0.10 -6.08  0.75  8.38 -0.99  0.06  1.30 -0.88
## fkgl                -1.38 -0.39  0.53 -0.50 -0.05  0.36 -0.20 -0.02  0.18 -2.11
## fre                 -0.10  0.06  0.22 -0.38  0.05  0.52 -0.64  0.05  0.82 -3.14
## hpoint              -0.18 -0.05  0.07 -7.52  0.98 10.45 -0.22  0.02  0.22 -0.08
## maentropy           -1.33 -0.36  0.53 -0.24  0.02  0.31 -0.33  0.00  0.41 -0.62
## entropy.v           -0.16 -0.07  0.12 -1.59  0.05  1.93 -3.18 -0.03  3.84 -0.40
## mamr                -0.92  0.82  2.67 -1.07 -0.04  0.87 -1.17  0.03  0.99 -0.12
## hapaxes             -0.10  0.04  0.20 -8.80 -0.82  6.33 -1.13  0.07  1.42 -0.77
## sentcount           -0.19  0.15  0.52 -7.13  0.94  9.91 -0.44  0.05  0.44 -1.22
## verbdist            -2.86 -0.79  1.10 -0.24  0.00  0.19 -0.32  0.10  0.49 -1.12
##                       PA4 upper   low   PA6 upper   low   PA5 upper   low   PA8
## VERBcomp             0.50  3.50  0.05  0.29  0.59 -0.21 -0.12 -0.04 -0.08  0.06
## literary             0.16  1.11 -0.55 -0.29 -0.06  0.01  0.14  0.33 -0.14 -0.03
## compoundVERBs       -0.28  1.04 -0.70 -0.31  0.00 -0.03  0.04  0.13 -0.22  0.16
## compoundVERBsdist.m -0.08  0.07 -0.21 -0.10  0.02 -0.20 -0.07  0.04 -0.30 -0.05
## compoundVERBsdist.v  0.01  0.16 -0.44 -0.19  0.02 -0.07  0.03  0.12 -0.14 -0.03
## passives            -0.22  0.86 -1.68 -0.82 -0.12  0.01  0.09  0.21 -0.32 -0.09
## predorder.m          0.17  1.14 -0.13  0.13  0.25 -0.16 -0.05  0.08 -0.55  0.00
## predorder.v          0.12  0.73 -0.10  0.05  0.14 -0.03  0.04  0.12 -0.22 -0.04
## obj                  0.91  5.95 -0.05  0.15  0.44  0.03  0.12  0.27 -0.36 -0.08
## predobjdist.m       -0.13  0.15 -0.12  0.05  0.18 -0.20 -0.06  0.04 -0.11  0.06
## predobjdist.v        0.04  0.46 -0.15 -0.01  0.11 -0.04  0.05  0.17 -0.15  0.03
## subj                -0.07  0.05 -0.25 -0.09  0.01 -0.02  0.06  0.16 -0.14  0.08
## predsubjdist.m       0.05  0.22 -0.11  0.14  0.38 -0.07  0.03  0.15 -0.51  0.18
## predsubjdist.v       0.13  0.80 -0.12  0.02  0.14 -0.02  0.08  0.21 -0.26 -0.06
## VERBfrac.m           0.03  0.11  0.06  0.35  0.72 -0.06 -0.01  0.06 -0.11  0.06
## VERBfrac.v          -0.18  0.61 -0.06  0.22  0.56 -0.13  0.02  0.17 -0.13  0.04
## NEGcount.m           0.17  1.28 -0.07  0.05  0.11  0.36  0.99  1.77 -0.11  0.00
## NEGcount.v           0.06  0.24 -0.12 -0.05  0.02  0.21  0.73  1.52 -0.04  0.04
## NEGfrac.m           -0.21  0.60  0.13  0.50  0.93  0.12  0.28  0.51 -0.44 -0.11
## NOUNcount.m         -0.03  0.13 -0.12 -0.02  0.05 -0.28 -0.14 -0.04 -0.09  0.02
## NOUNcount.v          0.05  0.17 -0.11 -0.02  0.12 -0.09 -0.02  0.08 -0.44 -0.11
## activity             0.25  1.52  0.11  0.49  0.97 -0.03  0.01  0.05 -0.38 -0.10
## cli                 -0.12  0.57 -0.09  0.10  0.23 -0.09  0.01  0.08 -1.54  0.79
## entropy             -0.13  0.53 -0.04  0.04  0.15  0.03  0.10  0.18 -0.23  0.19
## fkgl                 0.53  3.52 -0.59 -0.29 -0.06 -0.01  0.05  0.14 -0.40  0.18
## fre                 -0.46  1.87  0.04  0.20  0.48 -0.13 -0.06  0.01 -2.74 -0.63
## hpoint               0.02  0.13 -0.04  0.00  0.04 -0.05 -0.01  0.05 -0.22 -0.06
## maentropy           -0.10  0.35 -0.04  0.07  0.22  0.01  0.09  0.18 -0.49  0.23
## entropy.v            0.03  0.37 -0.10  0.05  0.28 -0.07  0.10  0.17 -0.23  0.03
## mamr                -0.05  0.05 -0.14  0.00  0.06 -0.13 -0.04  0.04 -0.21  0.15
## hapaxes             -0.12  0.46  0.00  0.08  0.16 -0.06  0.02  0.09 -0.18  0.13
## sentcount           -0.24  0.63  0.08  0.28  0.47 -0.15 -0.08 -0.04 -0.03  0.07
## verbdist            -0.22  0.60 -0.52 -0.17  0.05 -0.14 -0.07 -0.01 -0.37 -0.07
##                     upper    low   PA3 upper    low   PA9 upper
## VERBcomp             0.26  -2.52 -0.03  2.20  -1.53  0.00  1.37
## literary             0.12  -3.39 -0.03  2.93  -3.07  0.07  3.53
## compoundVERBs        0.65  -0.32  0.00  0.31  -1.02 -0.01  1.05
## compoundVERBsdist.m  0.19  -6.91  0.13  8.20  -0.42  0.08  0.73
## compoundVERBsdist.v  0.08  -1.63  0.08  2.05  -5.00  0.02  5.59
## passives             0.14  -0.89 -0.04  0.59  -0.89 -0.06  0.68
## predorder.m          0.36  -0.53 -0.16  0.29  -1.72 -0.26  1.32
## predorder.v          0.09  -2.98 -0.15  2.49  -9.76  0.03 11.02
## obj                  0.18  -0.11  0.04  0.17  -5.76 -0.06  4.96
## predobjdist.m        0.23 -11.40 -0.02 12.83  -5.98 -0.10  5.20
## predobjdist.v        0.27  -2.76 -0.05  3.12  -3.43  0.03  3.98
## subj                 0.33  -6.74  0.02  7.62 -10.39 -0.29  8.57
## predsubjdist.m       0.96 -19.79  0.31 23.01 -18.62 -0.29 15.93
## predsubjdist.v       0.09  -0.99 -0.02  1.14  -5.22 -0.01  5.80
## VERBfrac.m           0.24  -0.11  0.05  0.32  -0.14  0.02  0.18
## VERBfrac.v           0.17  -1.66  0.14  1.89  -4.03  0.20  5.11
## NEGcount.m           0.05  -1.02  0.02  1.23  -4.96 -0.04  4.46
## NEGcount.v           0.15  -1.10  0.06  1.19  -0.99  0.07  1.29
## NEGfrac.m            0.07  -2.91 -0.06  3.31  -3.10 -0.15  2.57
## NOUNcount.m          0.15  -0.15 -0.03  0.18  -2.37  0.04  2.83
## NOUNcount.v          0.19  -1.88  0.03  1.84  -7.04  0.26  8.62
## activity             0.12  -1.42  0.01  1.37  -0.11 -0.01  0.08
## cli                  3.66  -2.29 -0.01  2.49  -0.69 -0.01  0.66
## entropy              0.71  -7.49 -0.05  6.41 -15.37  0.39 18.29
## fkgl                 0.94  -0.40  0.02  0.36  -1.83 -0.01  1.58
## fre                  1.07  -1.32 -0.02  1.17  -2.74  0.07  3.24
## hpoint               0.08  -5.10  0.00  5.78  -3.13  0.03  3.62
## maentropy            1.13 -29.10 -0.42 24.61 -27.59  0.43 32.31
## entropy.v            0.19 -16.22  0.97 20.08 -13.27 -0.04 11.58
## mamr                 0.56  -6.34 -0.06  7.14 -11.33 -0.40  9.21
## hapaxes              0.49 -11.16 -0.09  9.66  -4.43  0.13  5.31
## sentcount            0.15  -2.90 -0.04  3.28  -5.56  0.03  6.41
## verbdist             0.14  -4.35 -0.05  4.87  -1.97 -0.13  1.51
##
##  Interfactor correlations and bootstrapped confidence intervals
##         lower estimate upper
## PA1-PA2 -0.22   0.1100  0.35
## PA1-PA7 -0.95  -0.6083  0.41
## PA1-PA4 -1.00  -0.2330  0.66
## PA1-PA6 -0.77   0.4138  0.44
## PA1-PA5 -0.66  -0.2598  0.36
## PA1-PA8 -0.62  -0.0827  0.35
## PA1-PA3 -0.44  -0.0208  0.35
## PA1-PA9 -0.44   0.0441  0.35
## PA2-PA7 -0.24   0.1582  0.54
## PA2-PA4 -0.44   0.3063  0.60
## PA2-PA6 -0.36  -0.2433  0.66
## PA2-PA5 -0.25   0.3067  0.51
## PA2-PA8 -0.27   0.1755  0.48
## PA2-PA3 -0.20   0.0358  0.30
## PA2-PA9 -0.27   0.1005  0.36
## PA7-PA4 -0.76   0.3949  0.82
## PA7-PA6 -0.61  -0.4245  0.81
## PA7-PA5 -0.42   0.2573  0.63
## PA7-PA8 -0.38   0.0582  0.57
## PA7-PA3 -0.40   0.2710  0.44
## PA7-PA9 -0.40  -0.0095  0.46
## PA4-PA6 -0.74  -0.4229  0.69
## PA4-PA5 -0.60   0.2611  0.59
## PA4-PA8 -0.45   0.2981  0.53
## PA4-PA3 -0.45  -0.0543  0.37
## PA4-PA9 -0.43   0.1124  0.37
## PA6-PA5 -0.37  -0.3032  0.44
## PA6-PA8 -0.38  -0.3153  0.41
## PA6-PA3 -0.41   0.0569  0.40
## PA6-PA9 -0.44  -0.0228  0.39
## PA5-PA8 -0.33   0.0288  0.35
## PA5-PA3 -0.36  -0.1509  0.32
## PA5-PA9 -0.37   0.1143  0.29
## PA8-PA3 -0.36  -0.1765  0.35
## PA8-PA9 -0.43   0.1008  0.30
## PA3-PA9 -0.33  -0.1162  0.34
```

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: 33 x 2
##    feat                 maxload
##    <chr>                  <dbl>
##  1 compoundVERBsdist.v    0.253
##  2 literary               0.291
##  3 predsubjdist.m         0.309
##  4 NOUNcount.v            0.389
##  5 maentropy              0.429
##  6 predsubjdist.v         0.468
##  7 NEGfrac.m             0.502
##  8 fkgl                  0.530
##  9 predorder.m           0.556
## 10 predobjdist.v         0.561
## # i 23 more rows
```

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

```
##            literary compoundVERBsdist.v          VERBfrac.v         NOUNcount.v
##           0.2362743           0.3257346           0.3667412           0.3677439
##       predobjdist.v       predobjdist.m           NEGfrac.m compoundVERBsdist.m
##           0.3791083           0.4080409           0.4230923           0.4600003
##      predsubjdist.v      predsubjdist.m            passives                subj
##           0.4646279           0.4765022           0.5535799           0.5626395
##         predorder.v          NEGcount.v            VERBcomp           maentropy
##           0.5762151           0.5849480           0.5929031           0.6018590
##         predorder.m                 obj                 cli       compoundVERBs
##           0.6936411           0.7040599           0.7199239           0.7220211
##             hapaxes                mamr            verbdist         NOUNcount.m
##           0.7361063           0.8048572           0.8091458           0.8127657
##             entropy           sentcount          VERBfrac.m           entropy.v
##           0.8566010           0.8834452           0.9008474           0.9154860
##            activity              hpoint                fkgl          NEGcount.m
##           0.9252350           0.9537280           0.9614465           0.9708204
##                 fre
##           0.9797952
```

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

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

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

```
##              hpoint           entropy.v         NOUNcount.m          NEGcount.m
##            1.016433            1.052208            1.067421            1.087144
##                 obj         predorder.v       predobjdist.m       predobjdist.v
##            1.168289            1.172894            1.191040            1.216418
##             hapaxes          NEGcount.v            passives            verbdist
##            1.220513            1.235699            1.243350            1.380477
## compoundVERBsdist.m           sentcount          VERBfrac.m                 cli
##            1.389810            1.412346            1.431277            1.526821
##                mamr       compoundVERBs      predsubjdist.v                subj
##            1.563896            1.646632            1.666644            1.725198
##             entropy            activity          VERBfrac.v                 fre
##            1.844008            2.055543            2.068799            2.192414
##           NEGfrac.m            literary            VERBcomp         NOUNcount.v
##            2.359770            2.525897            2.552161            2.692993
##         predorder.m                fkgl compoundVERBsdist.v           maentropy
##            2.720897            2.808915            3.522730            3.779264
##      predsubjdist.m
##            5.120289
```

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

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

Feature engineering

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

det(cor(data_engineered_1))
```

```
## [1] 1.306983e-07
```

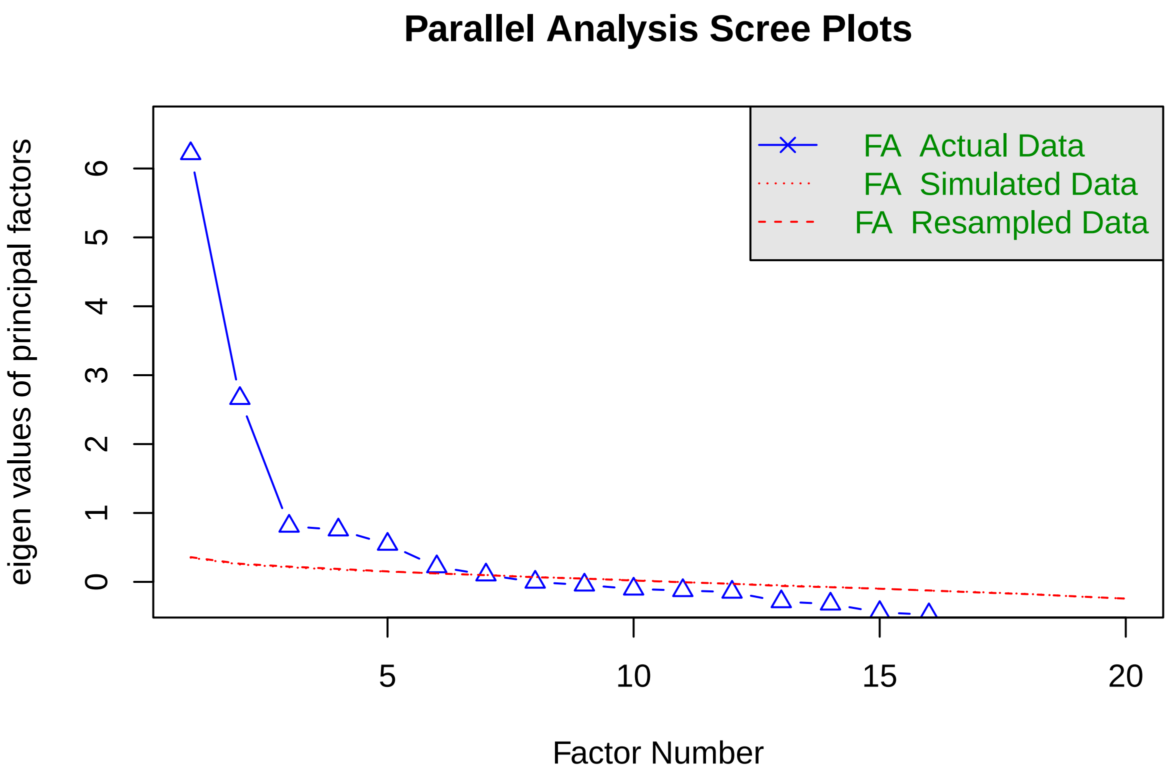
```
KMO(data_engineered_1)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = data_engineered_1)
## Overall MSA =  0.82
## MSA for each item =
##       VERBcomp compoundVERBs      passives   predorder.m   predorder.v
##           0.86          0.90          0.77          0.87          0.83
##            obj          subj    VERBfrac.m    NEGcount.m    NEGcount.v
##           0.56          0.94          0.88          0.72          0.67
##    NOUNcount.m      activity       entropy        hpoint     maentropy
##           0.91          0.89          0.68          0.70          0.56
##      entropy.v          mamr       hapaxes     sentcount      verbdist
##           0.37          0.91          0.77          0.73          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 =  6  and the number of components =  NA
```

Model

```
set.seed(42)

fa_2 <- fa(
  data_engineered_1,
  nfactors = 6,
  fm = "pa",
  rotate = "promax",
  oblique.scores = TRUE,
  scores = "tenBerge",
  n.iter = 100
)
```

```
## maximum iteration exceeded

## 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_2
```

```
## Factor Analysis with confidence intervals using method = fa(r = data_engineered_1, nfactors = 6, n.i
##     scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Factor Analysis using method =  pa
## Call: fa(r = data_engineered_1, nfactors = 6, n.iter = 100, rotate = "promax",
##     scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##                 PA1   PA2   PA3   PA4   PA5   PA6   h2     u2 com
## VERBcomp       0.52  0.05  0.09 -0.13 -0.18  0.45 0.60  0.40 2.4
## compoundVERBs  0.86  0.00  0.05  0.00  0.30 -0.05 0.61  0.39 1.3
## passives       0.07  0.02  0.07  0.05  0.75 -0.01 0.55  0.45 1.0
## predorder.m   -0.66 -0.06 -0.01 -0.14  0.13  0.31 0.61  0.39 1.6
## predorder.v   -0.43  0.09 -0.01  0.01  0.09  0.31 0.36  0.64 2.0
## obj           -0.01 -0.01  0.01  0.13 -0.03  0.72 0.57  0.43 1.1
## subj           0.75  0.12 -0.05 -0.02  0.20 -0.01 0.52  0.48 1.2
## VERBfrac.m     0.79 -0.04 -0.03 -0.03 -0.26  0.03 0.89  0.11 1.2
## NEGcount.m    -0.06 -0.08 -0.02  0.84  0.05  0.17 0.79  0.21 1.1
## NEGcount.v     0.16  0.06 -0.04  0.81  0.04  0.04 0.69  0.31 1.1
## NOUNcount.m   -0.93  0.03 -0.02 -0.13  0.00 -0.03 0.81  0.19 1.0
## activity       0.68 -0.05 -0.05  0.01 -0.40  0.17 0.89  0.11 1.8
## entropy       -0.08  0.81  0.27  0.10 -0.02 -0.04 0.76  0.24 1.3
## hpoint        -0.03  0.97 -0.05  0.00  0.07  0.07 0.97  0.03 1.0
## maentropy     -0.36  0.17  1.13  0.00 -0.10 -0.02 1.42 -0.42 1.3
## entropy.v     -0.13  0.07 -0.43  0.03 -0.10 -0.03 0.20  0.80 1.4
## mamr           0.87 -0.06 -0.08 -0.13  0.19  0.06 0.71  0.29 1.2
## hapaxes        0.00 -0.77  0.20  0.01 -0.13 -0.12 0.71  0.29 1.2
## sentcount      0.21  0.88 -0.02 -0.10 -0.16 -0.18 0.85  0.15 1.3
## verbdist      -0.78 -0.04 -0.07 -0.09  0.17 -0.13 0.76  0.24 1.2
##
##                         PA1  PA2  PA3  PA4  PA5  PA6
## SS loadings           5.83 3.02 1.61 1.51 1.23 1.08
## Proportion Var        0.29 0.15 0.08 0.08 0.06 0.05
## Cumulative Var        0.29 0.44 0.52 0.60 0.66 0.71
## Proportion Explained  0.41 0.21 0.11 0.11 0.09 0.08
## Cumulative Proportion 0.41 0.62 0.73 0.84 0.92 1.00
##
##  With factor correlations of
##       PA1   PA2   PA3   PA4   PA5   PA6
## PA1  1.00  0.09 -0.04 -0.20 -0.45 -0.07
## PA2  0.09  1.00 -0.05  0.33 -0.02  0.21
## PA3 -0.04 -0.05  1.00  0.16 -0.01 -0.08
## PA4 -0.20  0.33  0.16  1.00  0.21  0.20
## PA5 -0.45 -0.02 -0.01  0.21  1.00 -0.10
## PA6 -0.07  0.21 -0.08  0.20 -0.10  1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 6 factors are sufficient.
##
## df null model =  190  with the objective function =  15.85 with Chi Square =  11800.6
## df of  the model are 85  and the objective function was  1.24
##
## The root mean square of the residuals (RMSR) is  0.02
## The df corrected root mean square of the residuals is  0.04
##
## The harmonic n.obs is  753 with the empirical chi square  172.28  with prob <  7e-08
## The total n.obs was  753  with Likelihood Chi Square =  920.44  with prob <  2.2e-140
##
## Tucker Lewis Index of factoring reliability =  0.838
## RMSEA index =  0.114  and the 90 % confidence intervals are  0.108 0.121
## BIC =  357.39
## Fit based upon off diagonal values = 1
##  Coefficients and bootstrapped confidence intervals
##                  low   PA1 upper   low   PA2 upper   low   PA3 upper   low   PA4
## VERBcomp        0.44  0.52  0.58  0.01  0.05  0.10  0.04  0.09  0.13 -0.18 -0.13
## compoundVERBs   0.76  0.86  0.93 -0.05  0.00  0.06 -0.02  0.05  0.09 -0.07  0.00
## passives       -0.02  0.07  0.12 -0.02  0.02  0.06  0.00  0.07  0.12 -0.02  0.05
## predorder.m    -0.80 -0.66 -0.54 -0.11 -0.06 -0.02 -0.06 -0.01  0.03 -0.27 -0.14
## predorder.v    -0.52 -0.43 -0.37  0.03  0.09  0.15 -0.08 -0.01  0.06 -0.08  0.01
## obj            -0.07 -0.01  0.05 -0.05 -0.01  0.03 -0.04  0.01  0.05  0.04  0.13
## subj            0.69  0.75  0.81  0.07  0.12  0.17 -0.14 -0.05  0.03 -0.09 -0.02
## VERBfrac.m      0.73  0.79  0.84 -0.07 -0.04 -0.01 -0.07 -0.03  0.00 -0.10 -0.03
## NEGcount.m     -0.11 -0.06 -0.01 -0.12 -0.08 -0.03 -0.05 -0.02  0.02  0.74  0.84
## NEGcount.v      0.11  0.16  0.21  0.02  0.06  0.11 -0.08 -0.04  0.00  0.71  0.81
## NOUNcount.m    -0.96 -0.93 -0.87 -0.02  0.03  0.06 -0.05 -0.02  0.02 -0.20 -0.13
## activity        0.62  0.68  0.73 -0.08 -0.05 -0.02 -0.08 -0.05 -0.02 -0.04  0.01
## entropy        -0.13 -0.08 -0.04  0.78  0.81  0.85  0.22  0.27  0.34  0.04  0.10
## hpoint         -0.05 -0.03  0.00  0.94  0.97  0.99 -0.07 -0.05 -0.03 -0.03  0.00
## maentropy      -0.39 -0.36 -0.33  0.14  0.17  0.19  0.99  1.13  1.22 -0.02  0.00
## entropy.v      -0.20 -0.13 -0.06  0.01  0.07  0.14 -0.51 -0.43 -0.37 -0.05  0.03
## mamr            0.83  0.87  0.91 -0.12 -0.06 -0.01 -0.15 -0.08 -0.03 -0.20 -0.13
## hapaxes        -0.04  0.00  0.04 -0.80 -0.77 -0.74  0.17  0.20  0.25 -0.06  0.01
## sentcount       0.18  0.21  0.26  0.84  0.88  0.91 -0.05 -0.02  0.01 -0.18 -0.10
## verbdist       -0.84 -0.78 -0.73 -0.07 -0.04 -0.01 -0.13 -0.07 -0.03 -0.16 -0.09
##               upper   low   PA5 upper   low   PA6 upper
## VERBcomp      -0.05 -0.31 -0.18 -0.10  0.38  0.45  0.54
## compoundVERBs  0.10  0.15  0.30  0.42 -0.14 -0.05  0.01
## passives       0.15  0.65  0.75  0.86 -0.12 -0.01  0.05
## predorder.m    0.02 -0.01  0.13  0.24  0.18  0.31  0.45
## predorder.v    0.13 -0.06  0.09  0.23  0.16  0.31  0.46
## obj            0.25 -0.14 -0.03  0.06  0.61  0.72  0.86
## subj           0.06  0.09  0.20  0.32 -0.08 -0.01  0.06
## VERBfrac.m     0.03 -0.36 -0.26 -0.18 -0.04  0.03  0.10
## NEGcount.m     0.96  0.01  0.05  0.14  0.11  0.17  0.27
## NEGcount.v     0.94 -0.02  0.04  0.13 -0.02  0.04  0.14
## NOUNcount.m   -0.07 -0.06  0.00  0.05 -0.08 -0.03  0.03
## activity       0.06 -0.51 -0.40 -0.32  0.10  0.17  0.26
## entropy        0.16 -0.10 -0.02  0.04 -0.09 -0.04  0.01
## hpoint         0.04  0.03  0.07  0.12  0.04  0.07  0.10
## maentropy      0.05 -0.15 -0.10 -0.06 -0.06 -0.02  0.00
## entropy.v      0.11 -0.22 -0.10  0.00 -0.11 -0.03  0.05
## mamr          -0.04  0.08  0.19  0.30 -0.02  0.06  0.13
## hapaxes        0.06 -0.19 -0.13 -0.06 -0.17 -0.12 -0.07
## sentcount     -0.04 -0.22 -0.16 -0.09 -0.23 -0.18 -0.12
## verbdist      -0.03  0.07  0.17  0.31 -0.21 -0.13 -0.06
##
##  Interfactor correlations and bootstrapped confidence intervals
##           lower estimate upper
## PA1-PA2 -0.0049    0.086  0.18
## PA1-PA3 -0.4170   -0.040  0.16
## PA1-PA4 -0.4430   -0.204  0.19
## PA1-PA5 -0.7425   -0.447  0.10
## PA1-PA6 -0.3989   -0.069  0.17
## PA2-PA3 -0.2507   -0.054  0.48
## PA2-PA4 -0.2370    0.329  0.54
## PA2-PA5 -0.2042   -0.020  0.24
## PA2-PA6 -0.0418    0.205  0.34
## PA3-PA4  0.0277    0.161  0.26
## PA3-PA5 -0.1939   -0.013  0.31
## PA3-PA6 -0.2354   -0.084  0.32
## PA4-PA5 -0.1723    0.208  0.34
## PA4-PA6 -0.2207    0.204  0.36
## PA5-PA6 -0.2998   -0.096  0.17
```

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 entropy.v       0.428
##  2 predorder.v     0.434
##  3 VERBcomp        0.520
##  4 predorder.m     0.664
##  5 activity        0.679
##  6 obj             0.720
##  7 passives        0.751
##  8 subj            0.755
##  9 hapaxes         0.770
## 10 verbdist        0.783
## 11 VERBfrac.m      0.788
## 12 entropy         0.809
## 13 NEGcount.v      0.812
## 14 NEGcount.m      0.838
## 15 compoundVERBs   0.860
## 16 mamr            0.872
## 17 sentcount       0.877
## 18 NOUNcount.m     0.925
## 19 hpoint          0.965
## 20 maentropy       1.13
```

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

```
##     entropy.v   predorder.v          subj      passives           obj
##     0.1950318     0.3587653     0.5154365     0.5465891     0.5748423
##      VERBcomp compoundVERBs   predorder.m    NEGcount.v          mamr
##     0.6032729     0.6100970     0.6138020     0.6902068     0.7071752
##       hapaxes      verbdist       entropy    NEGcount.m   NOUNcount.m
##     0.7105518     0.7613928     0.7616281     0.7909550     0.8146737
##     sentcount      activity    VERBfrac.m        hpoint     maentropy
##     0.8548331     0.8871323     0.8876301     0.9701298     1.4159531
```

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

```
## [1] "predorder.v" "entropy.v"
```

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

```
##        hpoint      passives   NOUNcount.m           obj    NEGcount.v
##      1.029828      1.041563      1.043124      1.072235      1.104646
##    NEGcount.m          mamr          subj      verbdist    VERBfrac.m
##      1.118812      1.179233      1.206589      1.209112      1.234106
##       hapaxes compoundVERBs     maentropy       entropy     sentcount
##      1.246091      1.251573      1.263783      1.277751      1.315183
##     entropy.v   predorder.m      activity   predorder.v      VERBcomp
##      1.378335      1.621157      1.799766      2.023036      2.434814
```

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

```
## [1] "VERBcomp"    "predorder.v"
```

Feature engineering

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

det(cor(data_engineered_2))
```

```
## [1] 3.751325e-07
```

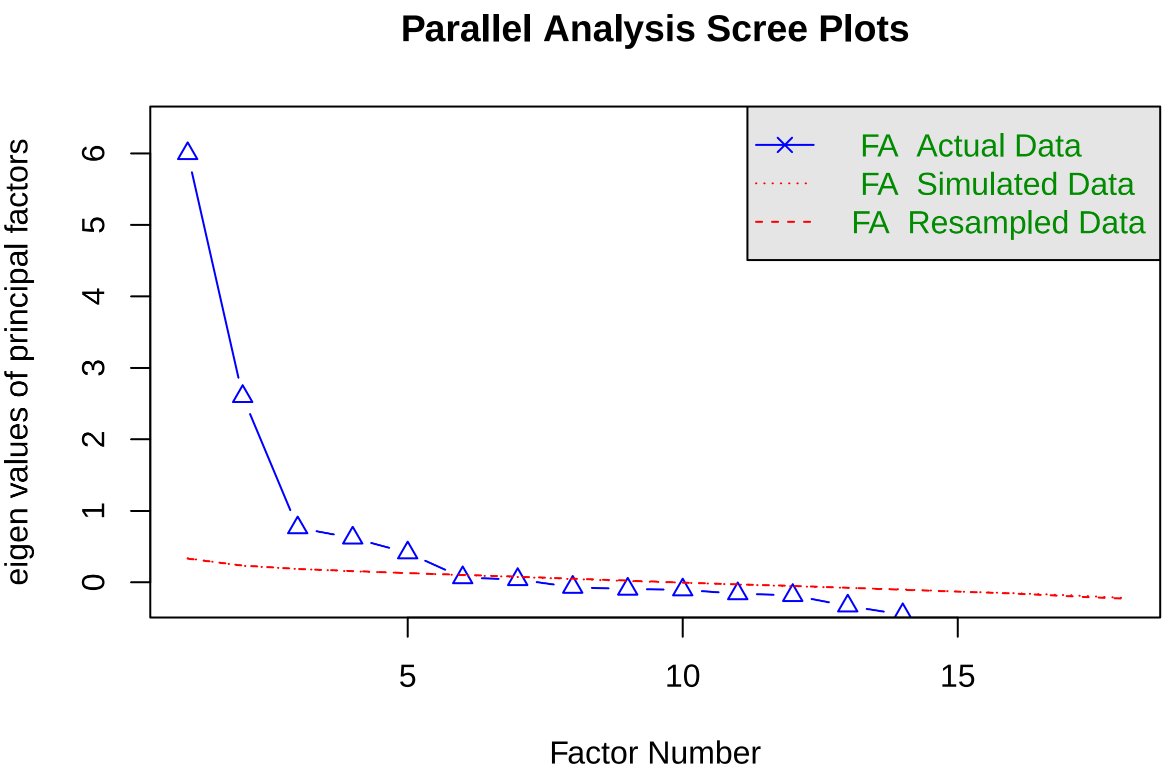
```
KMO(data_engineered_2)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = data_engineered_2)
## Overall MSA =  0.83
## MSA for each item = 
##      VERBcomp compoundVERBs      passives   predorder.m           obj
##          0.87          0.90          0.77          0.92          0.50
##          subj    VERBfrac.m    NEGcount.m    NEGcount.v   NOUNcount.m
##          0.93          0.88          0.70          0.67          0.91
##      activity       entropy        hpoint     maentropy          mamr
##          0.89          0.72          0.70          0.64          0.90
##       hapaxes     sentcount      verbdist
##          0.78          0.75          0.91
```

Third 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

```
set.seed(42)

fa_3 <- fa(
  data_engineered_2,
  nfactors = 5,
  fm = "pa",
  rotate = "promax",
  oblique.scores = TRUE,
  scores = "tenBerge",
  n.iter = 100
)
fa_3
```

```
## 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   PA4   PA3   PA5   h2    u2 com
## VERBcomp       0.31  0.05  0.55  0.07 -0.03 0.55 0.446 1.6
## compoundVERBs  0.86 -0.04 -0.22  0.05 -0.03 0.62 0.383 1.1
## passives       0.02  0.01 -0.60  0.20 -0.08 0.36 0.636 1.3
## predorder.m   -0.74 -0.01  0.00  0.03 -0.12 0.54 0.464 1.1
## obj           -0.24  0.05  0.41  0.43 -0.11 0.43 0.573 2.7
## subj           0.67  0.13 -0.10  0.03 -0.18 0.51 0.490 1.3
## VERBfrac.m     0.74 -0.06  0.35 -0.08 -0.01 0.90 0.103 1.5
## NEGcount.m     0.02 -0.09 -0.16  0.89  0.10 0.79 0.212 1.1
## NEGcount.v     0.25  0.06 -0.18  0.77  0.11 0.61 0.386 1.4
## NOUNcount.m   -0.90  0.07 -0.09 -0.14  0.00 0.82 0.175 1.1
## activity       0.58 -0.06  0.55 -0.01 -0.03 0.90 0.105 2.0
## entropy        0.03  0.73  0.01  0.09  0.51 0.92 0.082 1.8
## hpoint        -0.10  0.98 -0.01  0.05 -0.05 0.96 0.041 1.0
## maentropy     -0.15 -0.07  0.06  0.11  0.73 0.59 0.408 1.2
## mamr           0.73 -0.03 -0.02 -0.05 -0.29 0.71 0.290 1.3
## hapaxes        0.14 -0.87  0.05 -0.06  0.29 0.79 0.211 1.3
## sentcount      0.23  0.84  0.08 -0.21  0.08 0.81 0.194 1.3
## verbdist      -0.73  0.01 -0.32 -0.12 -0.09 0.77 0.228 1.5
##
##                        PA1  PA2  PA4  PA3  PA5
## SS loadings           5.07 2.97 1.73 1.69 1.11
## Proportion Var        0.28 0.17 0.10 0.09 0.06
## Cumulative Var        0.28 0.45 0.54 0.64 0.70
## Proportion Explained  0.40 0.24 0.14 0.13 0.09
## Cumulative Proportion 0.40 0.64 0.78 0.91 1.00
##
##  With factor correlations of
##       PA1  PA2   PA4   PA3   PA5
## PA1  1.00 0.16  0.39 -0.20 -0.20
## PA2  0.16 1.00  0.10  0.33  0.09
## PA4  0.39 0.10  1.00  0.04 -0.23
## PA3 -0.20 0.33  0.04  1.00  0.07
## PA5 -0.20 0.09 -0.23  0.07  1.00
##
## Mean item complexity =  1.4
## Test of the hypothesis that 5 factors are sufficient.
##
## df null model =  153  with the objective function =  14.8 with Chi Square =  11025.48
## df of  the model are 73  and the objective function was  1.33
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.04
##
## The harmonic n.obs is  753 with the empirical chi square  197.58  with prob <  2e-13
## The total n.obs was  753  with Likelihood Chi Square =  983.52  with prob <  1.7e-159
##
## Tucker Lewis Index of factoring reliability =  0.824
## RMSEA index =  0.129  and the 90 % confidence intervals are  0.122 0.136
## BIC =  499.97
## 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 0.92 0.93 0.92
## Multiple R square of scores with factors          0.95 0.98 0.84 0.87 0.84
## Minimum correlation of possible factor scores     0.89 0.96 0.69 0.73 0.69
##
##  Coefficients and bootstrapped confidence intervals
##                 low   PA1 upper   low   PA2 upper   low   PA4 upper   low   PA3
## VERBcomp       0.23  0.31  0.42  0.00  0.05  0.11  0.42  0.55  0.65  0.01  0.07
## compoundVERBs  0.79  0.86  0.94 -0.10 -0.04  0.01 -0.31 -0.22 -0.11 -0.01  0.05
## passives      -0.08  0.02  0.12 -0.04  0.01  0.07 -0.74 -0.60 -0.47  0.11  0.20
## predorder.m   -0.84 -0.74 -0.67 -0.05 -0.01  0.04 -0.09  0.00  0.08 -0.04  0.03
## obj           -0.36 -0.24 -0.12 -0.01  0.05  0.09  0.30  0.41  0.50  0.35  0.43
## subj           0.61  0.67  0.74  0.07  0.13  0.18 -0.17 -0.10 -0.02 -0.03  0.03
## VERBfrac.m     0.69  0.74  0.82 -0.09 -0.06 -0.03  0.29  0.35  0.40 -0.14 -0.08
## NEGcount.m    -0.05  0.02  0.07 -0.12 -0.09 -0.04 -0.21 -0.16 -0.09  0.80  0.89
## NEGcount.v     0.17  0.25  0.33  0.02  0.06  0.11 -0.25 -0.18 -0.10  0.70  0.77
## NOUNcount.m   -0.96 -0.90 -0.86  0.04  0.07  0.10 -0.14 -0.09 -0.05 -0.19 -0.14
## activity       0.53  0.58  0.68 -0.09 -0.06 -0.03  0.46  0.55  0.61 -0.05 -0.01
## entropy       -0.02  0.03  0.07  0.68  0.73  0.77 -0.03  0.01  0.06  0.05  0.09
## hpoint        -0.13 -0.10 -0.07  0.96  0.98  0.99 -0.05 -0.01  0.01  0.03  0.05
## maentropy     -0.21 -0.15 -0.11 -0.10 -0.07 -0.04  0.00  0.06  0.12  0.06  0.11
## mamr           0.67  0.73  0.81 -0.08 -0.03  0.02 -0.09 -0.02  0.05 -0.11 -0.05
## hapaxes        0.11  0.14  0.17 -0.89 -0.87 -0.84  0.01  0.05  0.09 -0.10 -0.06
## sentcount      0.19  0.23  0.28  0.80  0.84  0.88  0.04  0.08  0.13 -0.27 -0.21
## verbdist      -0.82 -0.73 -0.69 -0.02  0.01  0.04 -0.41 -0.32 -0.24 -0.20 -0.12
##               upper   low   PA5 upper
## VERBcomp       0.14 -0.11 -0.03  0.05
## compoundVERBs  0.12 -0.12 -0.03  0.04
## passives       0.30 -0.18 -0.08 -0.01
## predorder.m    0.11 -0.24 -0.12 -0.03
## obj            0.54 -0.21 -0.11 -0.01
## subj           0.09 -0.25 -0.18 -0.10
## VERBfrac.m    -0.03 -0.07 -0.01  0.04
## NEGcount.m     0.99  0.04  0.10  0.16
## NEGcount.v     0.85  0.04  0.11  0.20
## NOUNcount.m   -0.10 -0.04  0.00  0.04
## activity       0.03 -0.07 -0.03  0.02
## entropy        0.13  0.47  0.51  0.57
## hpoint         0.08 -0.07 -0.05 -0.02
## maentropy      0.16  0.65  0.73  0.82
## mamr           0.00 -0.39 -0.29 -0.22
## hapaxes       -0.03  0.24  0.29  0.33
## sentcount     -0.15  0.01  0.08  0.13
## verbdist      -0.04 -0.15 -0.09 -0.04
##
##  Interfactor correlations and bootstrapped confidence intervals
##          lower estimate upper
## PA1-PA2  0.083    0.161  0.25
## PA1-PA4 -0.664    0.392  0.62
## PA1-PA3 -0.596   -0.203  0.55
## PA1-PA5 -0.364   -0.204 -0.02
## PA2-PA4 -0.062    0.096  0.47
## PA2-PA3 -0.105    0.327  0.46
## PA2-PA5 -0.020    0.090  0.22
## PA4-PA3 -0.168    0.042  0.16
## PA4-PA5 -0.414   -0.234  0.36
## PA3-PA5 -0.315    0.070  0.35
```

Healthiness diagnostics

```
fa_3$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: 18 x 2
##    feat          maxload
##    <chr>           <dbl>
##  1 obj             0.435
##  2 VERBcomp        0.548
##  3 activity        0.584
##  4 passives        0.602
##  5 subj            0.672
##  6 maentropy       0.726
##  7 verbdist        0.730
##  8 entropy         0.731
##  9 mamr            0.732
## 10 predorder.m     0.740
## 11 VERBfrac.m      0.742
## 12 NEGcount.v      0.766
## 13 sentcount       0.836
## 14 compoundVERBs   0.855
## 15 hapaxes         0.869
## 16 NEGcount.m      0.893
## 17 NOUNcount.m     0.901
## 18 hpoint          0.976
```

```
fa_3$communality %>% sort()
```

```
##      passives           obj          subj   predorder.m      VERBcomp
##     0.3641885     0.4267584     0.5102071     0.5357288     0.5535377
##     maentropy    NEGcount.v compoundVERBs          mamr      verbdist
##     0.5920471     0.6138903     0.6169510     0.7100864     0.7719135
##    NEGcount.m       hapaxes     sentcount   NOUNcount.m      activity
##     0.7876946     0.7894769     0.8057570     0.8245139     0.8950871
##    VERBfrac.m       entropy        hpoint
##     0.8967049     0.9184282     0.9587591
```

```
fa_3$communality[fa_3$communality < 0.5] %>% names()
```

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

```
fa_3$complexity %>% sort()
```

```
##        hpoint   predorder.m   NOUNcount.m    NEGcount.m compoundVERBs
##      1.033724      1.061277      1.080563      1.108044      1.146598
##     maentropy      passives          subj       hapaxes          mamr
##      1.171025      1.263126      1.272920      1.300853      1.329067
##     sentcount    NEGcount.v    VERBfrac.m      verbdist      VERBcomp
##      1.331001      1.391961      1.465838      1.478578      1.640068
##       entropy      activity           obj
##      1.829554      2.017754      2.716342
```

```
fa_3$complexity[fa_3$complexity > 2] %>% names()
```

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

Feature engineering

```
data_engineered_3 <- data_engineered_2 %>%
  # remove low-communality features
  select(!c(
    passives,
    obj
  ))

det(cor(data_engineered_3))
```

```
## [1] 1.328369e-06
```

```
KMO(data_engineered_3)
```

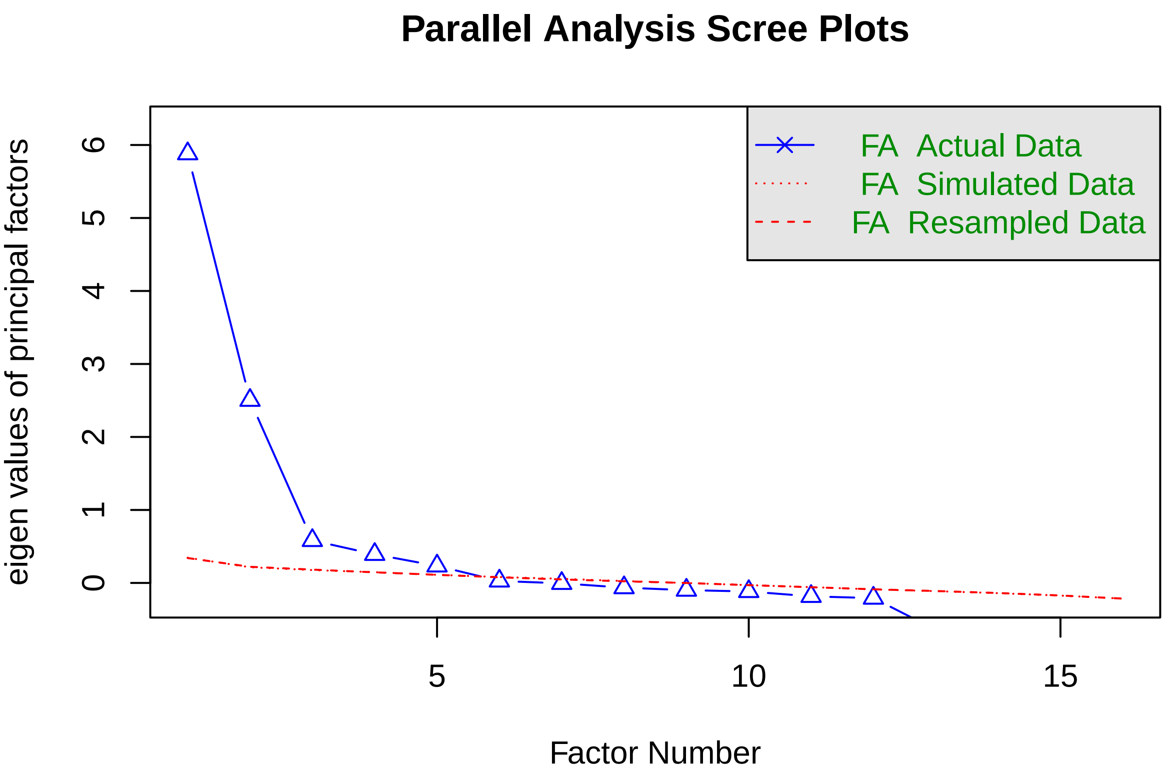
```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = data_engineered_3)
## Overall MSA =  0.84
## MSA for each item = 
##      VERBcomp compoundVERBs   predorder.m          subj    VERBfrac.m 
##          0.84          0.94          0.94          0.94          0.86 
##    NEGcount.m    NEGcount.v   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_3 %>% colnames()
```

Final FA

No. of vectors

```
fa.parallel(data_engineered_3, fm = "pa", fa = "fa", n.iter = 20)
```



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

Model

```r
set.seed(42)

fa_res <- fa(
  data_engineered_3,
  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_3, nfactors = 5, n.i
##     scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Factor Analysis using method =  pa
## Call: fa(r = data_engineered_3, 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       0.15  0.09  0.60  0.01 -0.01 0.52 0.482 1.2
## compoundVERBs  0.79 -0.06 -0.08  0.02  0.00 0.54 0.464 1.0
## predorder.m   -0.75  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 0.460 1.2
## VERBfrac.m     0.60 -0.06  0.44 -0.06 -0.03 0.90 0.098 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   -0.88  0.07 -0.09 -0.10 -0.02 0.83 0.166 1.1
## activity       0.39 -0.03  0.65  0.01 -0.06 0.90 0.095 1.6
## entropy        0.10  0.71 -0.06  0.01  0.55 0.95 0.054 1.9
## hpoint        -0.13  0.98  0.03  0.06 -0.05 0.96 0.041 1.1
## maentropy     -0.08 -0.11 -0.03  0.01  0.77 0.64 0.360 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 -0.08  0.29 0.77 0.229 1.3
## sentcount      0.21  0.80  0.09 -0.15  0.06 0.77 0.232 1.3
## verbdist      -0.69  0.00 -0.29 -0.07 -0.10 0.75 0.246 1.4
##
##                        PA1  PA2  PA5  PA3  PA4
## SS loadings           4.64 2.95 1.55 1.52 1.15
## Proportion Var        0.29 0.18 0.10 0.10 0.07
## Cumulative Var        0.29 0.47 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  0.61 -0.17 -0.26
## PA2  0.18 1.00  0.07  0.29  0.16
## PA5  0.61 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 =  13.53 with Chi Square =  10092.29
## df of  the model are 50  and the objective function was  0.75
##
## 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  753 with the empirical chi square  60.52  with prob <  0.15
## The total n.obs was  753  with Likelihood Chi Square =  559.19  with prob <  3.4e-87
##
## Tucker Lewis Index of factoring reliability =  0.877
## RMSEA index =  0.116  and the 90 % confidence intervals are  0.108 0.125
## BIC =  227.99
## 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     0.88 0.97 0.77 0.76 0.75
##
##  Coefficients and bootstrapped confidence intervals
##                 low   PA1 upper   low   PA2 upper   low   PA5 upper   low   PA3
## VERBcomp       0.03  0.15  0.29  0.05  0.09  0.13  0.42  0.60  0.80 -0.04  0.01
## compoundVERBs  0.72  0.79  0.85 -0.11 -0.06  0.00 -0.17 -0.08  0.03 -0.04  0.02
## predorder.m   -0.89 -0.75 -0.66 -0.03  0.02  0.07 -0.10  0.02  0.12 -0.05  0.03
## subj           0.68  0.75  0.80  0.06  0.11  0.17 -0.25 -0.16 -0.05 -0.05  0.00
## VERBfrac.m     0.54  0.60  0.69 -0.09 -0.06 -0.03  0.33  0.44  0.55 -0.10 -0.06
## NEGcount.m    -0.15 -0.11 -0.06 -0.09 -0.05 -0.02 -0.01  0.04  0.09  0.83  0.91
## NEGcount.v     0.11  0.17  0.23  0.04  0.07  0.13 -0.10 -0.03  0.03  0.69  0.80
## NOUNcount.m   -0.98 -0.88 -0.80  0.04  0.07  0.11 -0.19 -0.09 -0.01 -0.14 -0.10
## activity       0.30  0.39  0.51 -0.07 -0.03  0.00  0.49  0.65  0.82 -0.03  0.01
## entropy        0.05  0.10  0.13  0.68  0.71  0.75 -0.11 -0.06 -0.02 -0.02  0.01
## hpoint        -0.16 -0.13 -0.10  0.97  0.98  1.01 -0.01  0.03  0.07  0.03  0.06
## maentropy     -0.13 -0.08 -0.03 -0.14 -0.11 -0.08 -0.08 -0.03  0.02 -0.02  0.01
## mamr           0.64  0.74  0.84 -0.08 -0.04  0.00 -0.14 -0.02  0.11 -0.10 -0.05
## hapaxes        0.12  0.18  0.22 -0.91 -0.88 -0.86 -0.07 -0.01  0.04 -0.12 -0.08
## sentcount      0.15  0.21  0.29  0.77  0.80  0.84  0.03  0.09  0.16 -0.18 -0.15
## verbdist      -0.77 -0.69 -0.63 -0.03  0.00  0.03 -0.45 -0.29 -0.16 -0.12 -0.07
##               upper   low   PA4 upper
## VERBcomp       0.05 -0.08 -0.01  0.06
## compoundVERBs  0.08 -0.06  0.00  0.05
## predorder.m    0.13 -0.18 -0.12 -0.05
## subj           0.07 -0.21 -0.14 -0.06
## VERBfrac.m    -0.03 -0.07 -0.03  0.01
## NEGcount.m     1.03 -0.03  0.00  0.04
## NEGcount.v     0.88 -0.02  0.02  0.07
## NOUNcount.m   -0.06 -0.06 -0.02  0.01
## activity       0.04 -0.10 -0.06 -0.02
## entropy        0.05  0.49  0.55  0.60
## hpoint         0.08 -0.08 -0.05 -0.02
## maentropy      0.06  0.72  0.77  0.84
## mamr           0.01 -0.33 -0.26 -0.20
## hapaxes       -0.04  0.23  0.29  0.33
## sentcount     -0.11  0.01  0.06  0.10
## verbdist      -0.01 -0.14 -0.10 -0.06
##
##  Interfactor correlations and bootstrapped confidence intervals
##             lower estimate upper
## PA1-PA2  0.10127    0.184  0.27
## PA1-PA5 -0.64382    0.608  0.90
## PA1-PA3 -0.68628   -0.165  0.87
## PA1-PA4 -0.62777   -0.259  0.36
## PA2-PA5  0.00069    0.071  0.42
## PA2-PA3 -0.03696    0.289  0.37
## PA2-PA4 -0.05209    0.162  0.28
## PA5-PA3 -0.46608   -0.173  0.32
## PA5-PA4 -0.39174   -0.152  0.44
## PA3-PA4 -0.38667    0.282  0.43
```

Healthiness diagnostics

```
fa_res$loadings[] %>%
  as_tibble() %>%
  mutate(feat = colnames(data_engineered_3)) %>%
  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        0.599
##  2 VERBfrac.m      0.601
##  3 activity        0.655
##  4 verbdist        0.691
##  5 entropy         0.711
##  6 mamr            0.737
##  7 subj            0.746
##  8 predorder.m     0.754
##  9 maentropy       0.774
## 10 compoundVERBs   0.787
## 11 NEGcount.v      0.799
## 12 sentcount       0.801
## 13 hapaxes         0.885
## 14 NOUNcount.m     0.885
## 15 NEGcount.m      0.907
## 16 hpoint          0.985
```

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

```
##   predorder.m      VERBcomp compoundVERBs          subj     maentropy
##     0.5179923     0.5182886     0.5358740     0.5402714     0.6400470
##    NEGcount.v          mamr      verbdist     sentcount       hapaxes
##     0.6778257     0.7129269     0.7536391     0.7678487     0.7713750
##    NEGcount.m   NOUNcount.m    VERBfrac.m      activity       entropy
##     0.8300184     0.8343470     0.9022079     0.9045390     0.9460138
##        hpoint
##     0.9591754
```

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

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

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

```
## compoundVERBs    NEGcount.m        hpoint   predorder.m     maentropy
##      1.030601      1.038853      1.050821      1.058590      1.063675
##   NOUNcount.m    NEGcount.v      VERBcomp          subj     sentcount
##      1.064972      1.111958      1.182944      1.214355      1.256174
##          mamr       hapaxes      verbdist      activity    VERBfrac.m
##      1.261994      1.313925      1.409391      1.646873      1.884085
##       entropy
##      1.943688
```

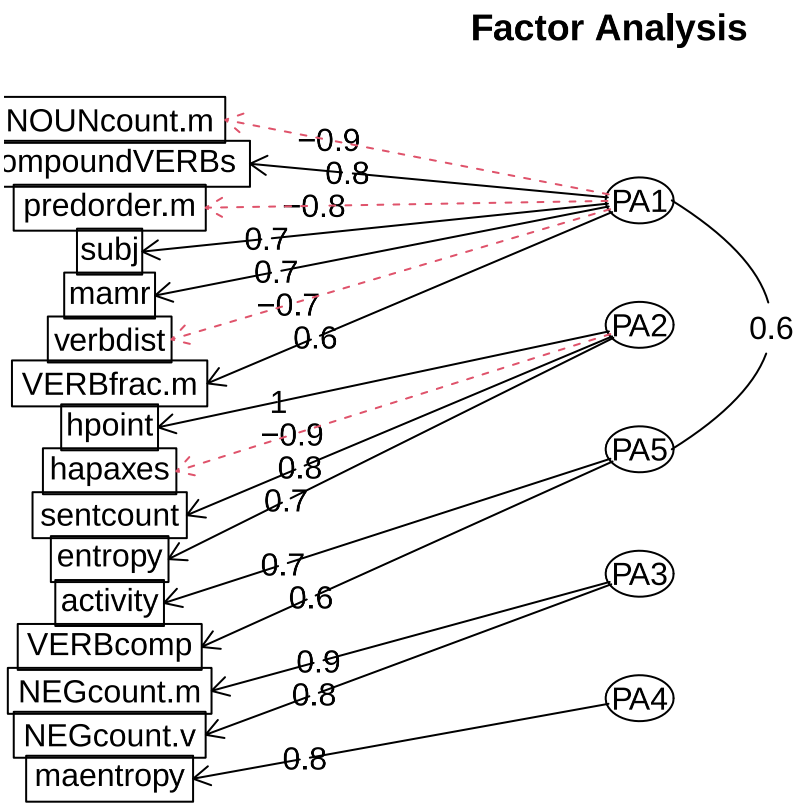
```
fa_res$complexity[fa_res$complexity > 2] %>% names()
```

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

Loadings

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

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



```r
fa_res$loadings
```

```
##
## Loadings:
##                PA1    PA2    PA5    PA3    PA4   
## VERBcomp       0.154         0.599               
## compoundVERBs  0.787                             
## predorder.m   -0.754                      -0.121 
## subj           0.746  0.115 -0.158        -0.140 
## VERBfrac.m     0.601         0.437               
## NEGcount.m    -0.109                0.907        
## NEGcount.v     0.169                0.799        
## NOUNcount.m   -0.885               -0.103        
## activity       0.385         0.655               
## entropy               0.711                0.547 
## hpoint        -0.134  0.985                      
## maentropy            -0.109                0.774 
## mamr           0.737                      -0.259 
## hapaxes        0.176 -0.885                0.286 
## sentcount      0.214  0.801        -0.149        
## verbdist      -0.691        -0.289        -0.100 
##
##                  PA1   PA2   PA5   PA3   PA4
## SS loadings    4.233 2.956 1.121 1.517 1.102
## Proportion Var 0.265 0.185 0.070 0.095 0.069
## Cumulative Var 0.265 0.449 0.519 0.614 0.683
```

```r
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(strng = 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 strng
## NOUNcount.m    -0.885 0.885 *****
## compoundVERBs   0.787 0.787 *****
## predorder.m    -0.754 0.754 *****
## subj            0.746 0.746 *****
## mamr            0.737 0.737 *****
## verbdist       -0.691 0.691 ****
## VERBfrac.m      0.601 0.601 ***
## activity        0.385 0.385 *
## sentcount       0.214 0.214
## hapaxes         0.176 0.176
## NEGcount.v      0.169 0.169
## VERBcomp        0.154 0.154
## hpoint         -0.134 0.134
## NEGcount.m     -0.109 0.109
##
##
## ----- PA2 -----
##           loading abs_l strng
## hpoint      0.985 0.985 *****
## hapaxes    -0.885 0.885 *****
## sentcount   0.801 0.801 *****
## entropy     0.711 0.711 *****
## subj        0.115 0.115
## maentropy  -0.109 0.109
##
##
## ----- PA5 -----
##             loading abs_l strng
## activity     0.655 0.655 ****
## VERBcomp     0.599 0.599 ***
## VERBfrac.m   0.437 0.437 *
## verbdist    -0.289 0.289
## subj        -0.158 0.158
##
##
## ----- PA3 -----
##              loading abs_l strng
## NEGcount.m    0.907 0.907 *****
## NEGcount.v    0.799 0.799 *****
## sentcount    -0.149 0.149
## NOUNcount.m  -0.103 0.103
##
##
## ----- PA4 -----
##              loading abs_l strng
## maentropy     0.774 0.774 *****
## entropy       0.547 0.547 **
## hapaxes       0.286 0.286
## mamr         -0.259 0.259
## subj         -0.140 0.140
## predorder.m  -0.121 0.121
## verbdist     -0.100 0.100
```

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

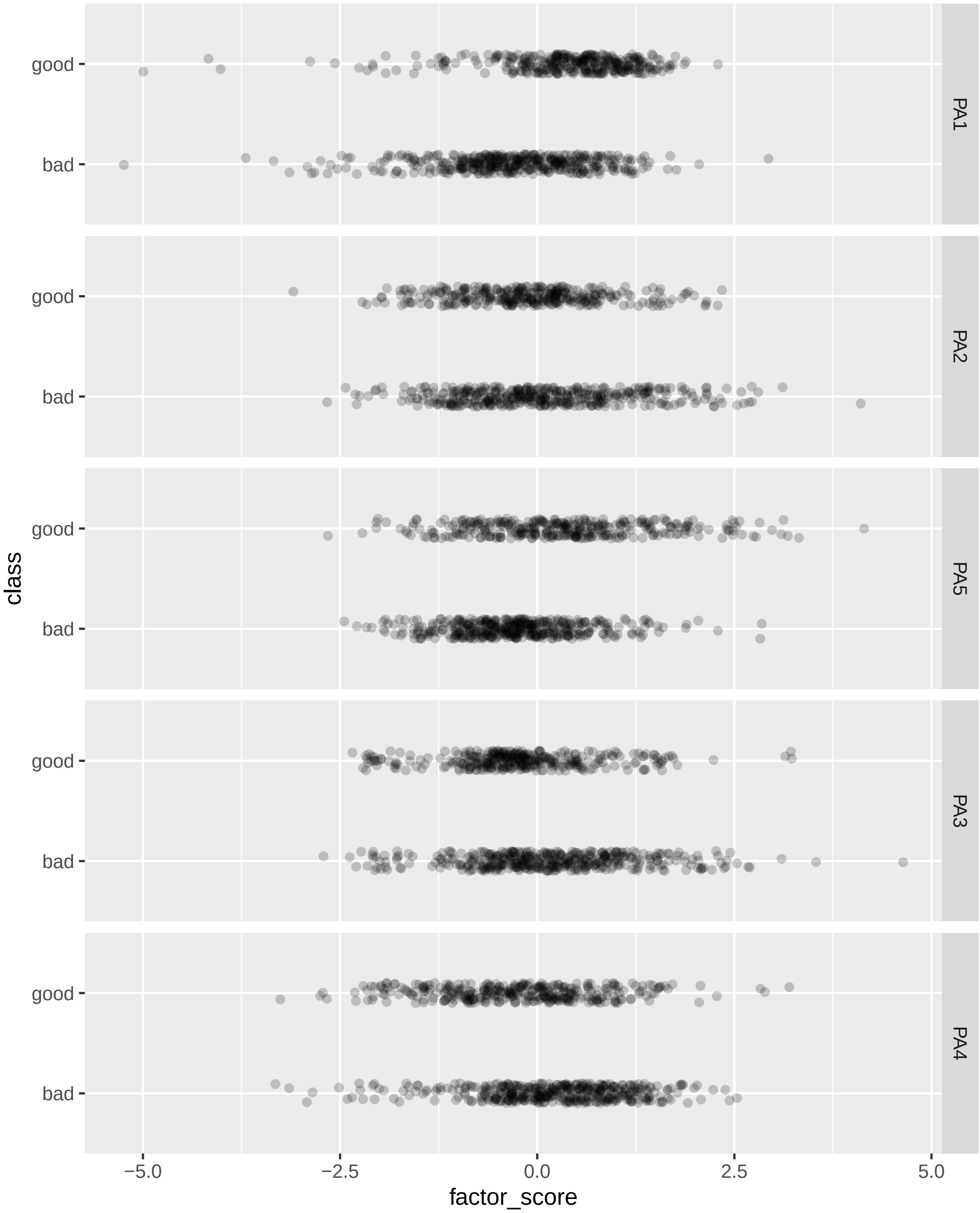
```
fa_res$uniquenesses %>% round(3)
```

```
##      VERBcomp compoundVERBs   predorder.m          subj    VERBfrac.m
##         0.482         0.464         0.482         0.460         0.098
##    NEGcount.m    NEGcount.v   NOUNcount.m      activity       entropy
##         0.170         0.322         0.166         0.095         0.054
##        hpoint     maentropy          mamr       hapaxes     sentcount
##         0.041         0.360         0.287         0.229         0.232
##      verbdist
##         0.246
```

Distributions over factors

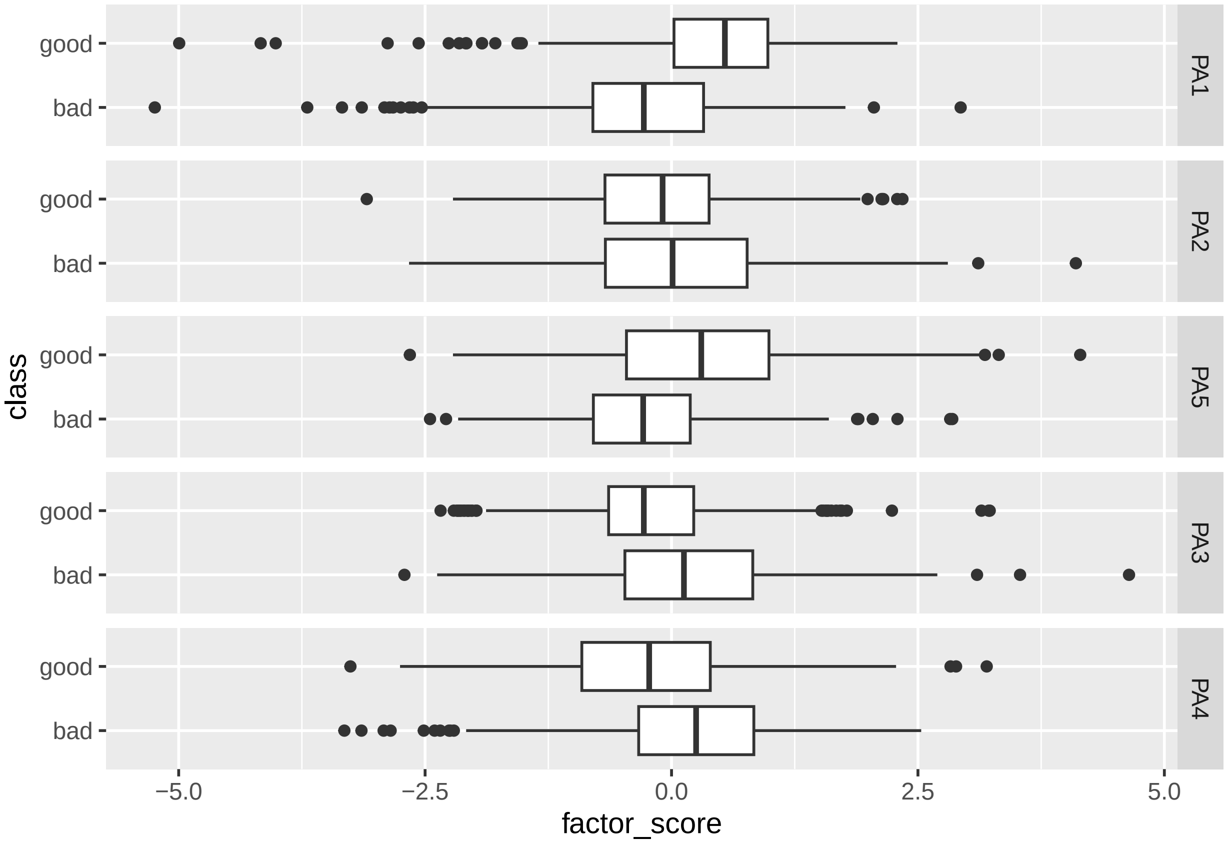
```
analyze_distributions <- function(data_factors_long, variable) {
  plot <- data_factors_long %>%
    ggplot(aes(x = factor_score, y = !!sym(variable))) +
    geom_boxplot() +
    facet_grid(factor ~ .)
  print(plot)

  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)
```

```r
  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

```r
analyze_distributions(data_factors_long, "class")
```



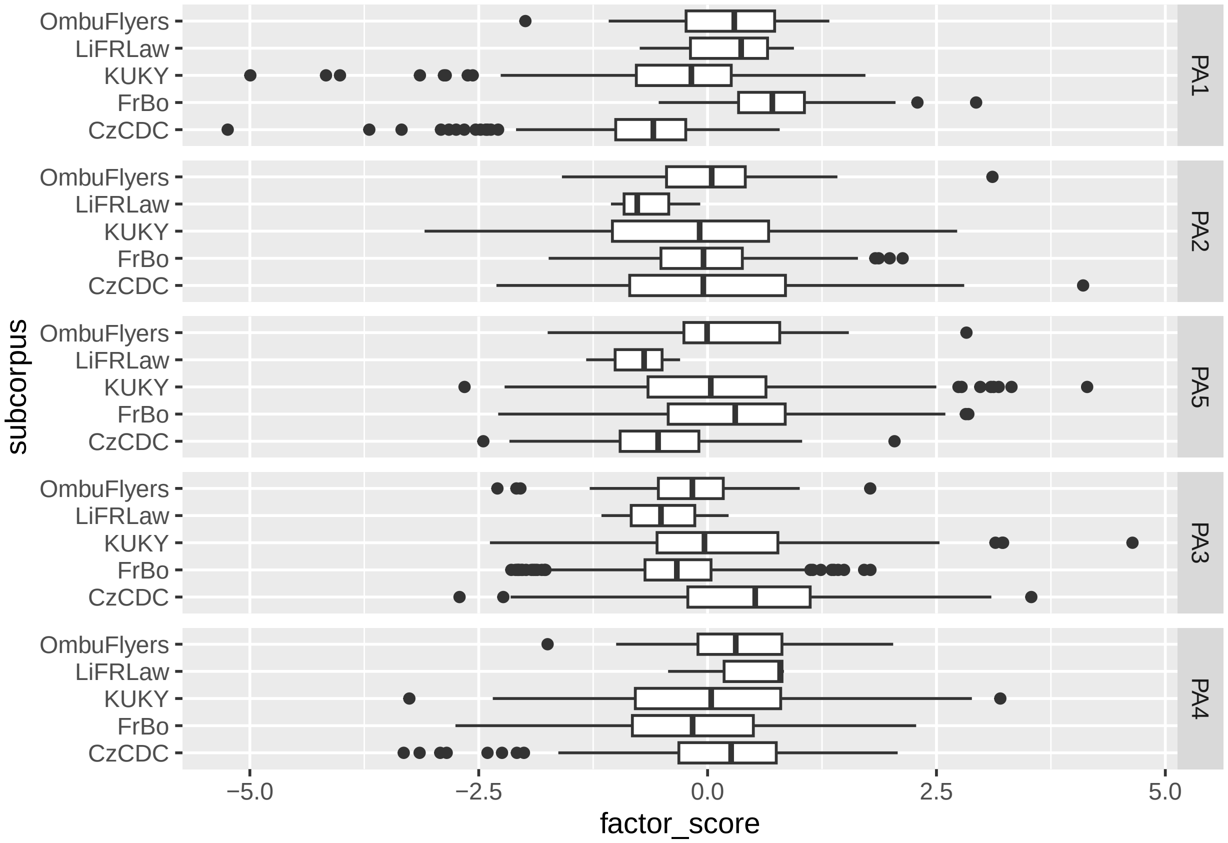
```
##
## Test for the significance of differences in class over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 123.8025, df = 1, p-value = 0
##
##
##                              Comparison of x by group
##                                    (Bonferroni)
## Col Mean-|
## Row Mean |        bad
## ---------+-----------
##     good |  -11.12665
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.165
##
## Test for the significance of differences in class over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 4.419, df = 1, p-value = 0.04
```

```
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |        bad
## ---------+-----------
##     good |   2.102148
##          |    0.0355*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.00588
##
## Test for the significance of differences in class over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 66.6336, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |        bad
## ---------+-----------
##     good |  -8.162938
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0886
##
## Test for the significance of differences in class over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 31.6013, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |        bad
## ---------+-----------
##     good |   5.621501
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.042
##
```

```
## Test for the significance of differences in class over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 42.0062, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |        bad
## ---------+-----------
##     good |   6.481219
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0559
##
##   factor kruskal_p epsilon2
## 1    PA1  9.31e-29  0.16500
## 2    PA2  3.55e-02  0.00588
## 3    PA5  3.27e-16  0.08860
## 4    PA3  1.89e-08  0.04200
## 5    PA4  9.10e-11  0.05590
##
## 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

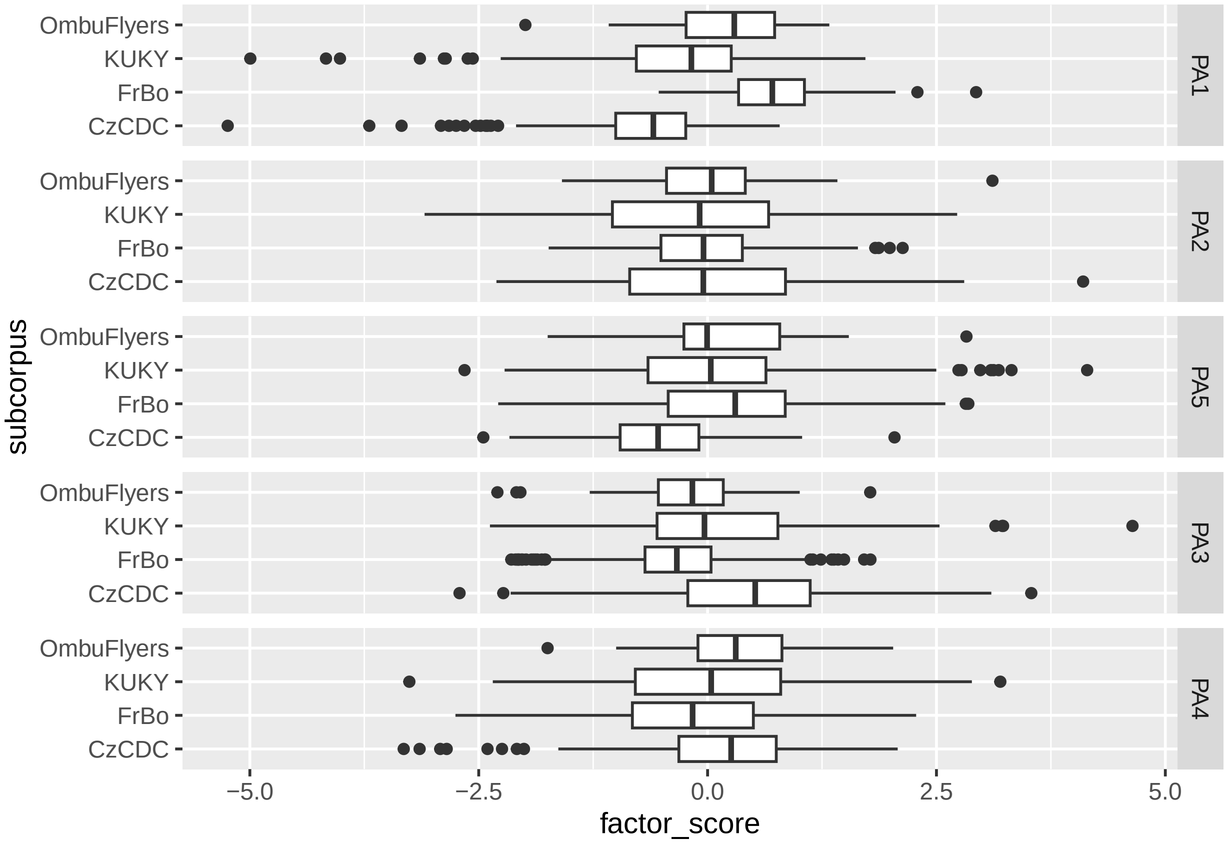
```
analyze_distributions(data_factors_long, "subcorpus")
```



```
##
## Test for the significance of differences in subcorpus over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 363.6725, df = 4, p-value = 0
##
##
##                              Comparison of x by group
##                                    (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY    LiFRLaw
## ---------+--------------------------------------------
##     FrBo |  -18.01448
##          |    0.0000*
##          |
##     KUKY |  -4.417524   12.77327
##          |    0.0001*    0.0000*
##          |
##  LiFRLaw |  -1.694035   1.078915  -0.937742
##          |     0.9026     1.0000     1.0000
##          |
## OmbuFlye |  -5.812922   3.410791  -3.297513  -0.065698
##          |    0.0000*    0.0065*    0.0098*     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.484
##
## Test for the significance of differences in subcorpus over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 4.8193, df = 4, p-value = 0.31
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY    LiFRLaw
## ---------+--------------------------------------------
##     FrBo |   0.700290
##          |     1.0000
##          |
##     KUKY |   1.626804   1.081512
##          |     1.0000     1.0000
##          |
##  LiFRLaw |   1.398422   1.293557   1.119433
##          |     1.0000     1.0000     1.0000
##          |
## OmbuFlye |  -0.239750  -0.609837  -1.150319  -1.426276
##          |     1.0000     1.0000     1.0000     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.00641
##
## Test for the significance of differences in subcorpus over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 113.196, df = 4, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY    LiFRLaw
## ---------+--------------------------------------------
##     FrBo |  -10.26540
##          |    0.0000*
##          |
##     KUKY |  -6.794022   2.640555
##          |    0.0000*     0.0828
##          |
##  LiFRLaw |   0.552478   2.135959   1.713697
##          |     1.0000     0.3268     0.8658
##          |
## OmbuFlye |  -4.889762   0.327255  -1.047952  -1.972511
##          |    0.0000*     1.0000     1.0000     0.4855
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.151
##
## Test for the significance of differences in subcorpus over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 98.4022, df = 4, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY    LiFRLaw
## ---------+--------------------------------------------
##     FrBo |   9.807405
##          |    0.0000*
##          |
##     KUKY |   4.673215  -4.494058
##          |    0.0000*    0.0001*
##          |
##  LiFRLaw |   1.847412   0.339803   1.047310
##          |     0.6469     1.0000     1.0000
##          |
## OmbuFlye |   3.734895  -1.272545   1.089876  -0.693637
##          |    0.0019*     1.0000     1.0000     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.131
##
## Test for the significance of differences in subcorpus over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 24.2893, df = 4, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY    LiFRLaw
## ---------+--------------------------------------------
##     FrBo |   4.183277
##          |    0.0003*
##          |
##     KUKY |   2.017488  -1.890702
##          |     0.4364     0.5866
##          |
##  LiFRLaw |  -0.421322  -1.067042  -0.765989
##          |     1.0000     1.0000     1.0000
##          |
## OmbuFlye |  -1.117115  -3.320080  -2.240934   0.080223
##          |     1.0000    0.0090*     0.2503     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0323
##
##   factor kruskal_p epsilon2
## 1    PA1  1.96e-77  0.48400
## 2    PA2  3.06e-01  0.00641
## 3    PA5  1.51e-23  0.15100
## 4    PA3  2.15e-20  0.13100
## 5    PA4  6.99e-05  0.03230
##
## 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"
)
```



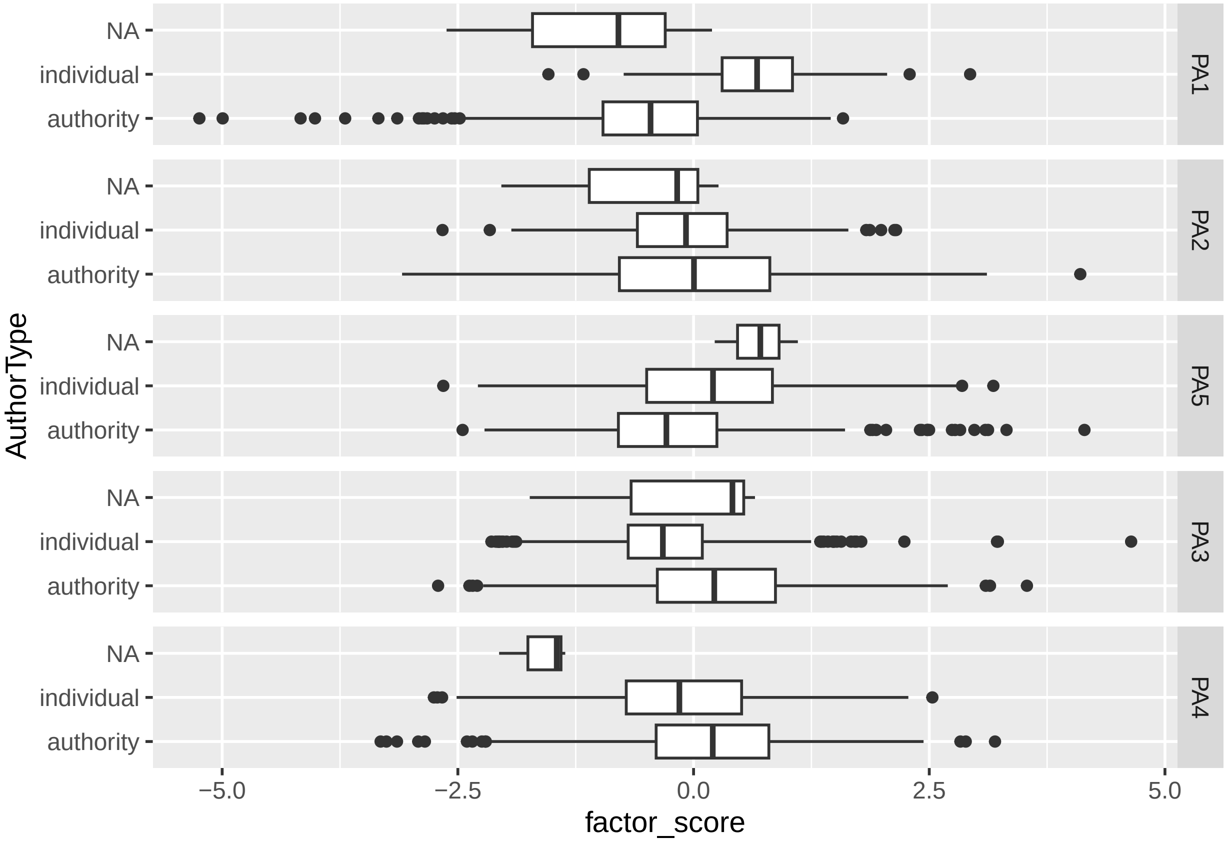
```
##
## Test for the significance of differences in subcorpus over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 363.4485, df = 3, p-value = 0
##
##
##                             Comparison of x by group
##                                   (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY
## ---------+---------------------------------
##     FrBo |  -18.01168
##          |    0.0000*
##          |
##     KUKY |  -4.418766   12.76920
##          |    0.0001*    0.0000*
##          |
## OmbuFlye |  -5.809810   3.412525  -3.293725
##          |    0.0000*    0.0039*    0.0059*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.485
##
```

```
## Test for the significance of differences in subcorpus over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 3.14, df = 3, p-value = 0.37
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY
## ---------+---------------------------------
##     FrBo |   0.716784
##          |     1.0000
##          |
##     KUKY |   1.628476   1.067244
##          |     0.6205     1.0000
##          |
## OmbuFlye |  -0.230922  -0.609367  -1.142487
##          |     1.0000     1.0000     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.00419
##
## Test for the significance of differences in subcorpus over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 110.831, df = 3, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY
## ---------+---------------------------------
##     FrBo |  -10.27209
##          |    0.0000*
##          |
##     KUKY |  -6.801608   2.638849
##          |    0.0000*    0.0499*
##          |
## OmbuFlye |  -4.888725   0.331795  -1.042668
##          |    0.0000*     1.0000     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.148
##
## Test for the significance of differences in subcorpus over PA3 :
##
```

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 97.4744, df = 3, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY
## ---------+---------------------------------
##     FrBo |   9.807962
##          |    0.0000*
##          |
##     KUKY |   4.671423  -4.496545
##          |    0.0000*    0.0000*
##          |
## OmbuFlye |   3.734958  -1.272770   1.090943
##          |    0.0011*     1.0000     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.13
##
## Test for the significance of differences in subcorpus over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 23.7336, df = 3, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY
## ---------+---------------------------------
##     FrBo |   4.185520
##          |    0.0002*
##          |
##     KUKY |   2.020834  -1.889262
##          |     0.2598     0.3531
##          |
## OmbuFlye |  -1.117131  -3.321264  -2.242826
##          |     1.0000    0.0054*     0.1494
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0317
##
##   factor kruskal_p epsilon2
## 1    PA1  1.83e-78  0.48500
## 2    PA2  3.71e-01  0.00419
## 3    PA5  7.27e-24  0.14800
## 4    PA3  5.43e-21  0.13000
## 5    PA4  2.84e-05  0.03170
##
## 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")
```

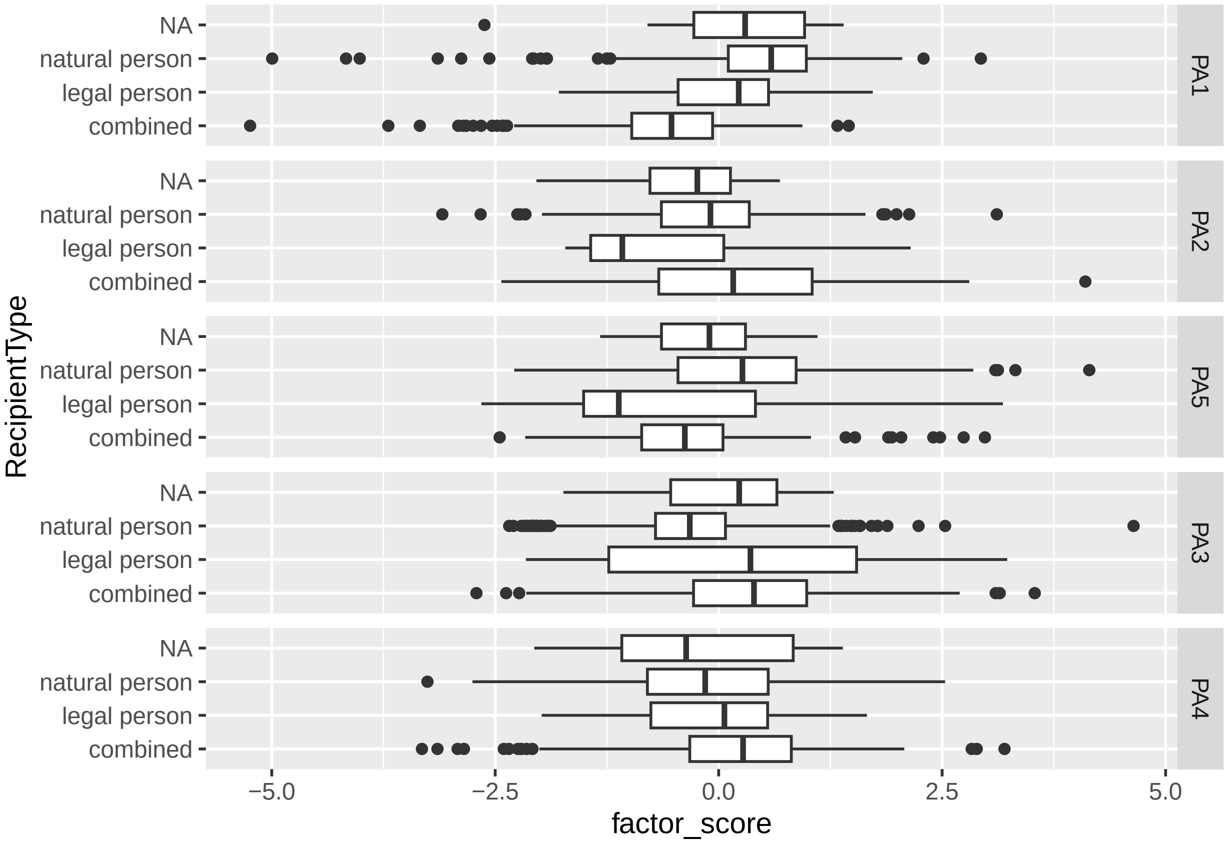


```
##
## Test for the significance of differences in AuthorType over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 337.0782, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## ---------+-----------
## individu |  -18.35969
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.448
##
## Test for the significance of differences in AuthorType over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 1.7573, df = 1, p-value = 0.18
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## ---------+-----------
## individu |   1.325641
##          |     0.1850
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.00234
##
## Test for the significance of differences in AuthorType over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 44.2164, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## ---------+-----------
## individu |  -6.649544
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0588
##
## Test for the significance of differences in AuthorType over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 59.6091, df = 1, p-value = 0
##
##
```

```
##                             Comparison of x by group
##                                   (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## ---------+-----------
## individu |   7.720691
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0793
##
## Test for the significance of differences in AuthorType over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 17.4734, df = 1, p-value = 0
##
##
##                             Comparison of x by group
##                                   (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## ---------+-----------
## individu |   4.180114
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0232
##
##   factor kruskal_p epsilon2
## 1    PA1  2.76e-75  0.44800
## 2    PA2  1.85e-01  0.00234
## 3    PA5  2.94e-11  0.05880
## 4    PA3  1.16e-14  0.07930
## 5    PA4  2.91e-05  0.02320
##
## 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")
```



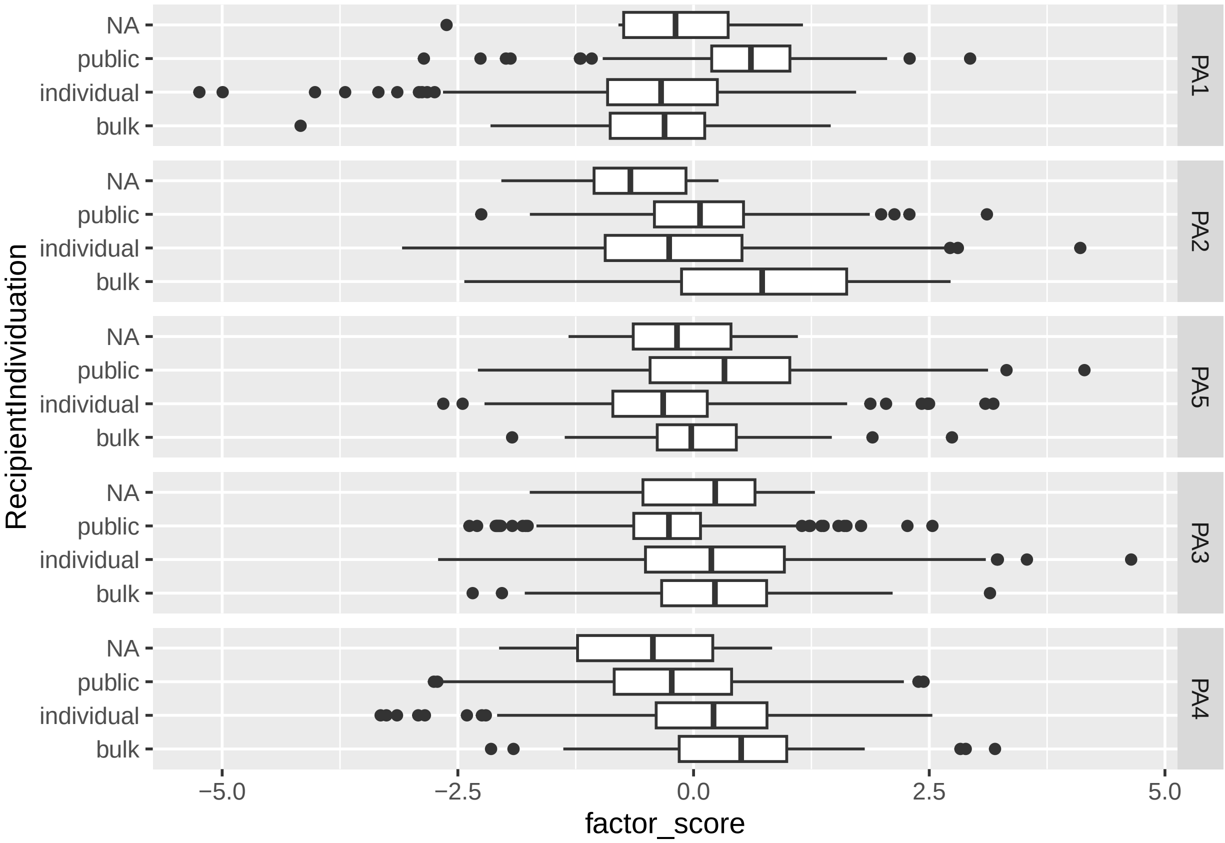
```
##
## Test for the significance of differences in RecipientType over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 272.2069, df = 2, p-value = 0
##
##
##                              Comparison of x by group
##                                    (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## ---------+----------------------
## legal pe |  -3.549157
##          |    0.0012*
##          |
## natural  |  -16.49704  -2.236450
##          |    0.0000*     0.0760
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.362
##
## Test for the significance of differences in RecipientType over PA2 :
##
##   Kruskal-Wallis rank sum test
```

```
##
## data: x and group
## Kruskal-Wallis chi-squared = 23.3932, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## ---------+----------------------
## legal pe |   3.898839
##          |    0.0003*
##          |
## natural  |   3.588398  -2.669800
##          |    0.0010*    0.0228*
##
## 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 = 94.5004, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## ---------+----------------------
## legal pe |   0.168203
##          |     1.0000
##          |
## natural  |  -9.486890  -3.516105
##          |    0.0000*    0.0013*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.126
##
## Test for the significance of differences in RecipientType over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 100.2001, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## ---------+----------------------
## legal pe |   1.264011
##          |     0.6187
##          |
## natural  |   9.981062   2.244718
##          |    0.0000*     0.0744
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.133
##
## Test for the significance of differences in RecipientType over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 21.2278, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## ---------+----------------------
## legal pe |   1.245845
##          |     0.6385
##          |
## natural  |   4.595708   0.363476
##          |    0.0000*     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0282
##
##   factor kruskal_p epsilon2
## 1    PA1  7.78e-60   0.3620
## 2    PA2  8.32e-06   0.0311
## 3    PA5  3.02e-21   0.1260
## 4    PA3  1.75e-22   0.1330
## 5    PA4  2.46e-05   0.0282
##
## 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

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

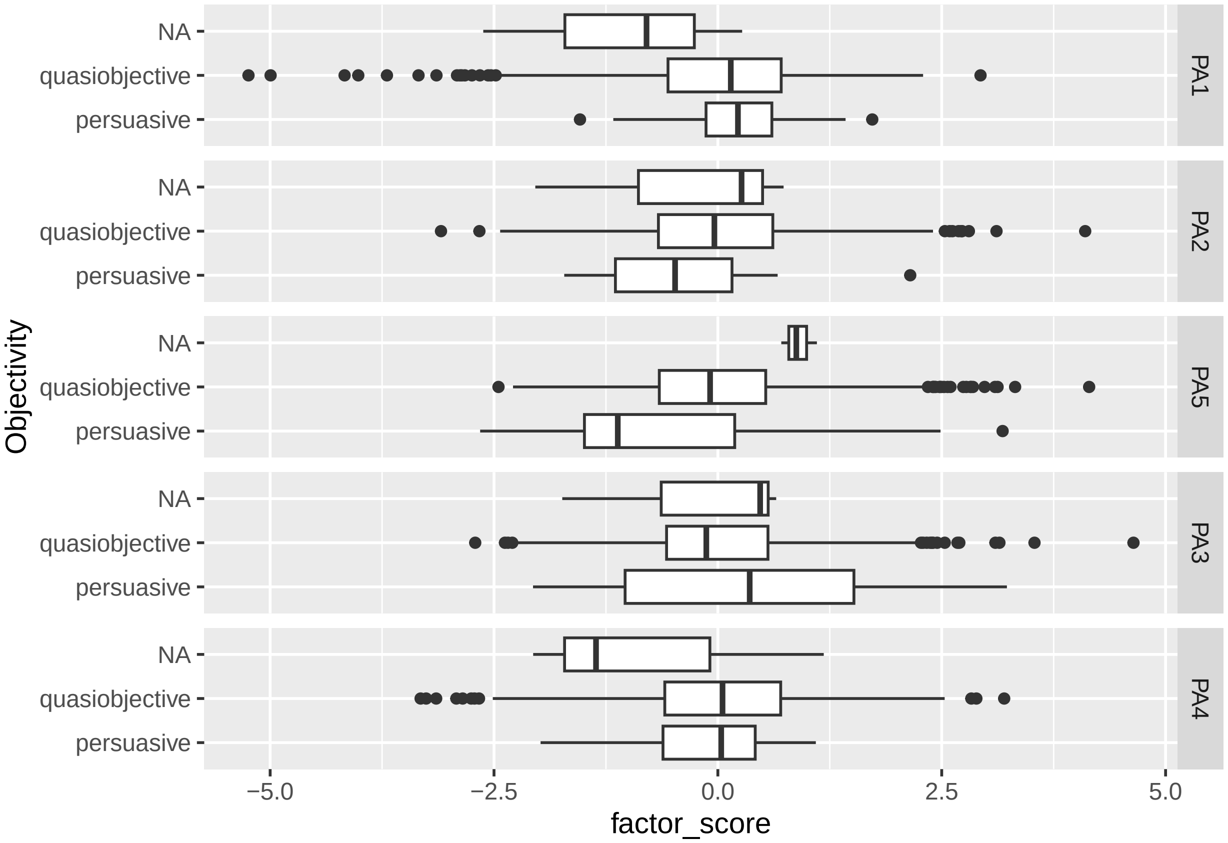


```
##
## Test for the significance of differences in RecipientIndividuation over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 210.8299, df = 2, p-value = 0
##
##
##                              Comparison of x by group
##                                    (Bonferroni)
## Col Mean-|
## Row Mean |       bulk   individu
## ---------+----------------------
## individu |  -0.733862
##          |     1.0000
##          |
##   public |  -8.700181  -13.73072
##          |    0.0000*    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.28
##
## Test for the significance of differences in RecipientIndividuation over PA2 :
##
##   Kruskal-Wallis rank sum test
```

```
##
## data: x and group
## Kruskal-Wallis chi-squared = 39.5755, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |       bulk   individu
## ---------+----------------------
## individu |   5.842865
##          |    0.0000*
##          |
##   public |   3.547872  -3.858839
##          |    0.0012*    0.0003*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0526
##
## Test for the significance of differences in RecipientIndividuation over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 74.4251, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |       bulk   individu
## ---------+----------------------
## individu |   2.925602
##          |    0.0103*
##          |
##   public |  -2.100389  -8.608604
##          |     0.1071    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.099
##
## Test for the significance of differences in RecipientIndividuation over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 45.165, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |       bulk   individu
## ---------+----------------------
## individu |   0.592664
##          |     1.0000
##          |
##   public |   4.226967   6.268197
##          |    0.0001*    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0601
##
## Test for the significance of differences in RecipientIndividuation over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 38.5192, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |       bulk   individu
## ---------+----------------------
## individu |   1.746288
##          |     0.2423
##          |
##   public |   4.772185   5.238890
##          |    0.0000*    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0512
##
##   factor kruskal_p epsilon2
## 1    PA1  1.66e-46   0.2800
## 2    PA2  2.55e-09   0.0526
## 3    PA5  6.90e-17   0.0990
## 4    PA3  1.56e-10   0.0601
## 5    PA4  4.32e-09   0.0512
##
## 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")
```



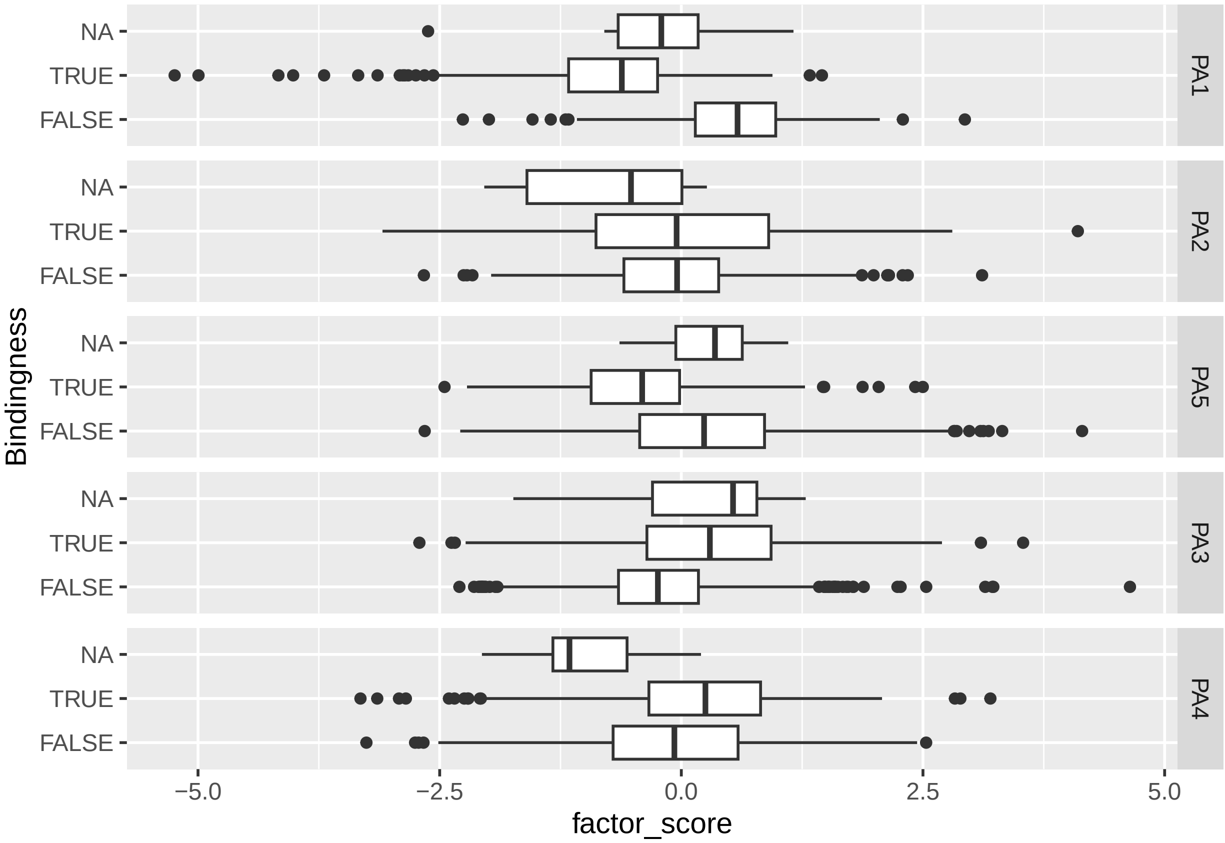
```
##
## Test for the significance of differences in Objectivity over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.5005, df = 1, p-value = 0.48
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## ---------+-----------
## quasiobj |   0.707484
##          |     0.4793
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.000666
##
## Test for the significance of differences in Objectivity over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 5.4329, df = 1, p-value = 0.02
```

```
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## ---------+-----------
## quasiobj |  -2.330868
##          |    0.0198*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.00722
##
## Test for the significance of differences in Objectivity over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 5.8552, df = 1, p-value = 0.02
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## ---------+-----------
## quasiobj |  -2.419750
##          |    0.0155*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.00779
##
## Test for the significance of differences in Objectivity over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.5816, df = 1, p-value = 0.45
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## ---------+-----------
## quasiobj |   0.762653
##          |     0.4457
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.000773
##
```

```
## Test for the significance of differences in Objectivity over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.3865, df = 1, p-value = 0.53
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## ---------+-----------
## quasiobj |  -0.621667
##          |     0.5342
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.000514
##
##   factor kruskal_p epsilon2
## 1    PA1    0.4793 0.000666
## 2    PA2    0.0198 0.007220
## 3    PA5    0.0155 0.007790
## 4    PA3    0.4457 0.000773
## 5    PA4    0.5342 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 = 352.8483, df = 1, p-value = 0
##
##
##                              Comparison of x by group
##                                    (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## ---------+-----------
##     TRUE |   18.78425
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.469
##
## Test for the significance of differences in Bindingness over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.8546, df = 1, p-value = 0.36
```

```
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## ---------+-----------
##     TRUE |  -0.924432
##          |     0.3553
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.00114
##
## Test for the significance of differences in Bindingness over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 99.1434, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## ---------+-----------
##     TRUE |   9.957078
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.132
##
## Test for the significance of differences in Bindingness over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 51.7954, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## ---------+-----------
##     TRUE |  -7.196901
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0689
##
```

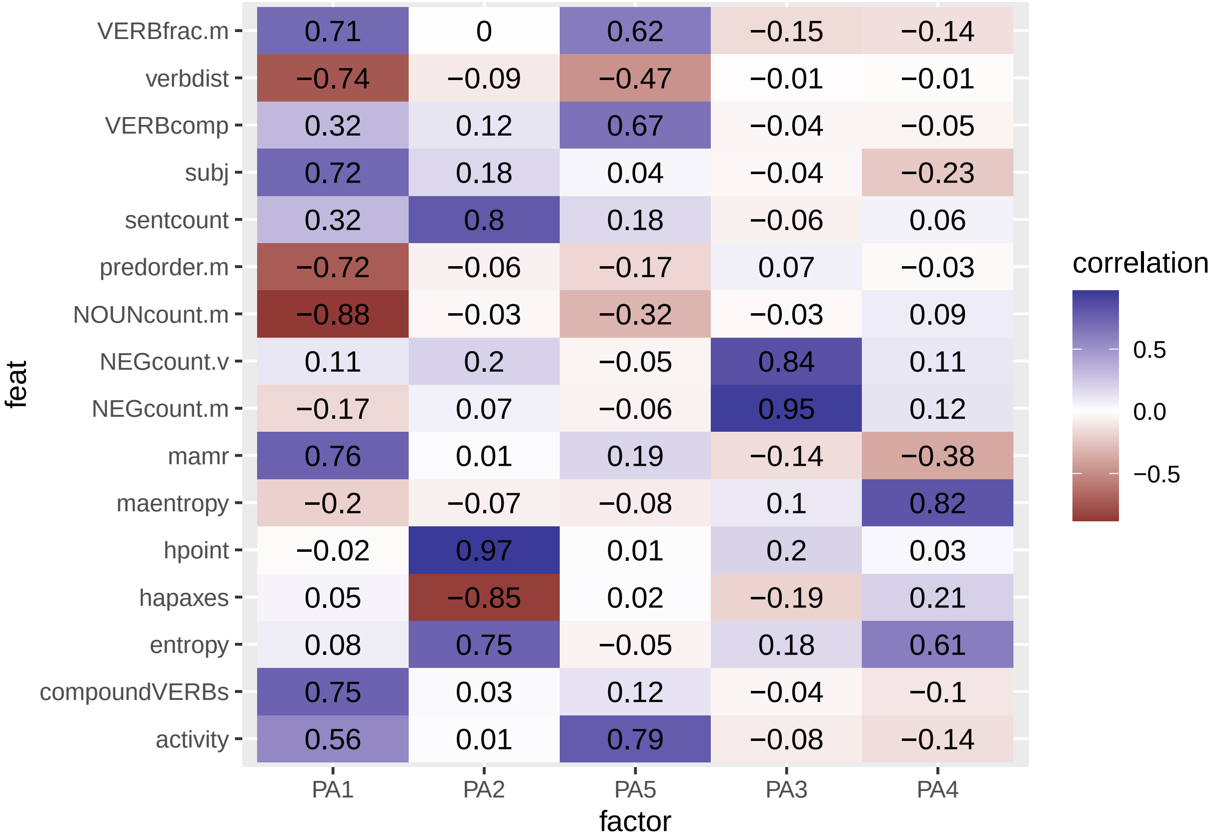
```
## Test for the significance of differences in Bindingness over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 16.5311, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## ---------+-----------
##     TRUE |  -4.065847
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.022
##
##   factor kruskal_p epsilon2
## 1    PA1  1.02e-78  0.46900
## 2    PA2  3.55e-01  0.00114
## 3    PA5  2.35e-23  0.13200
## 4    PA3  6.16e-13  0.06890
## 5    PA4  4.79e-05  0.02200
##
## 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

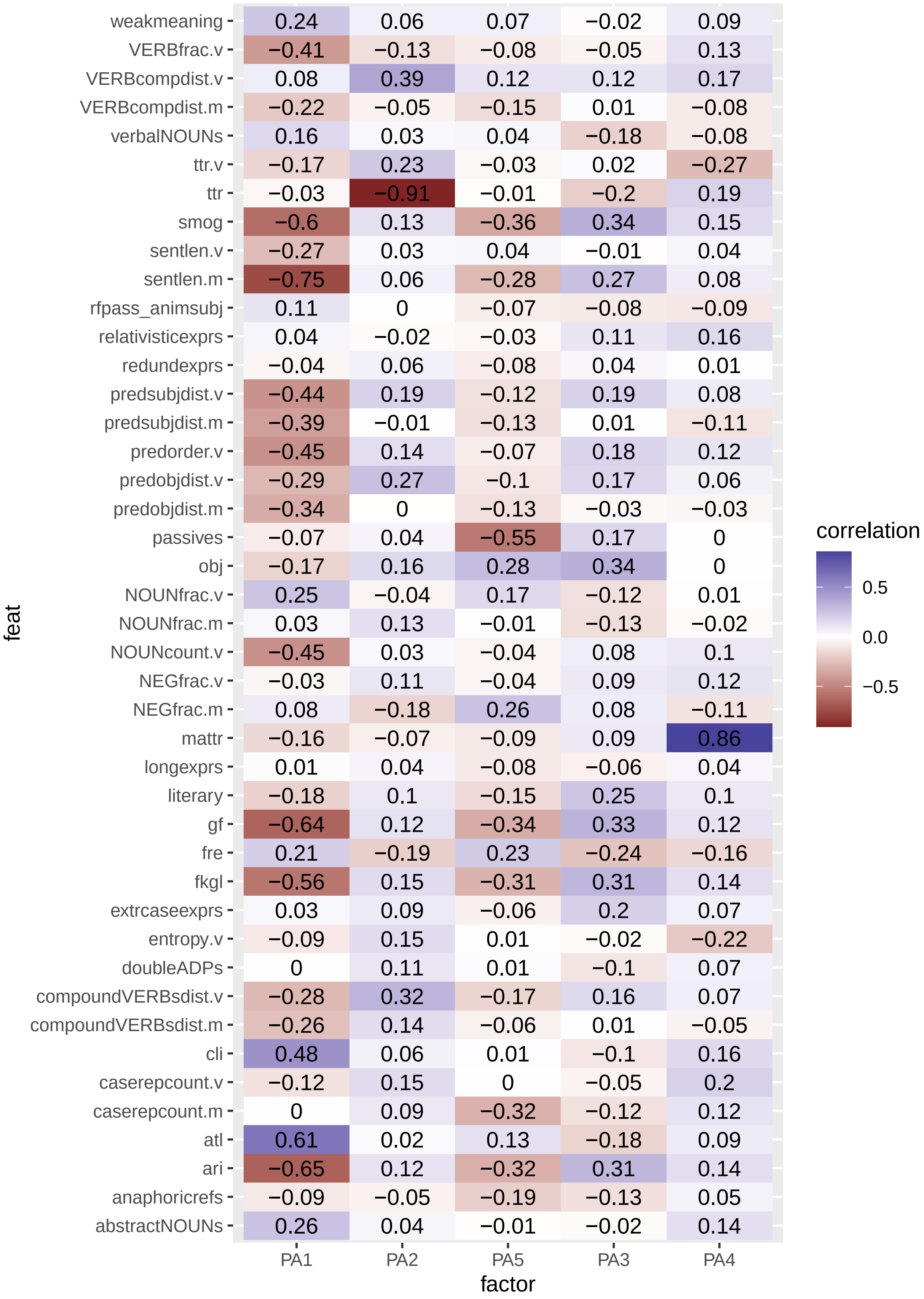
```
data_factors_longer <- data_factors_long %>%
  pivot_longer(
    abstractNOUNs:verbdist,
    names_to = "feat", values_to = "feat_value"
  )

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
factor
PA1
PA2
PA5
PA3
PA4
correlation
0.5
0.0
−0.5
weakmeaning 0.24 0.06 0.07 −0.02 0.09
VERBfrac.v −0.41 −0.13 −0.08 −0.05 0.13
VERBcompdist.v 0.08 0.39 0.12 0.12 0.17
VERBcompdist.m −0.22 −0.05 −0.15 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 −0.27
ttr −0.03 −0.91 −0.01 −0.2 0.19
smog −0.6 0.13 −0.36 0.34 0.15
sentlen.v −0.27 0.03 0.04 −0.01 0.04
sentlen.m −0.75 0.06 −0.28 0.27 0.08
rfpass_animsubj 0.11 0 −0.07 −0.08 −0.09
relativisticexprs 0.04 −0.02 −0.03 0.11 0.16
redundexprs −0.04 0.06 −0.08 0.04 0.01
predsubjdist.v −0.44 0.19 −0.12 0.19 0.08
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 0.17 0.06
predobjdist.m −0.34 0 −0.13 −0.03 −0.03
passives −0.07 0.04 −0.55 0.17 0
obj −0.17 0.16 0.28 0.34 0
NOUNfrac.v 0.25 −0.04 0.17 −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 0.11 −0.04 0.09 0.12
NEGfrac.m 0.08 −0.18 0.26 0.08 −0.11
mattr −0.16 −0.07 −0.09 0.09 0.86
longexprs 0.01 0.04 −0.08 −0.06 0.04
literary −0.18 0.1 −0.15 0.25 0.1
gf −0.64 0.12 −0.34 0.33 0.12
fre 0.21 −0.19 0.23 −0.24 −0.16
fkgl −0.56 0.15 −0.31 0.31 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
compoundVERBsdist.v −0.28 0.32 −0.17 0.16 0.07
compoundVERBsdist.m −0.26 0.14 −0.06 0.01 −0.05
cli 0.48 0.06 0.01 −0.1 0.16
caserepcount.v −0.12 0.15 0 −0.05 0.2
caserepcount.m 0 0.09 −0.32 −0.12 0.12
atl 0.61 0.02 0.13 −0.18 0.09
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