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

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

```
## # A tibble: 0 x 79
## # i 79 variables: KUK_ID <chr>, FileName <chr>, FileFormat <fct>,
## #   subcorpus <fct>, SourceID <chr>, DocumentVersion <fct>,
## #   ParentDocumentID <chr>, LegalActType <fct>, Objectivity <fct>,
## #   Bindingness <lgl>, AuthorType <fct>, RecipientType <fct>,
## #   RecipientIndividuation <fct>, Anonymized <fct>, Recipient Type <chr>,
## #   class <fct>, RuleAbstractNouns <dbl>, RuleAnaphoricReferences <dbl>,
## #   RuleCaseRepetition.max_repetition_count <dbl>, ...
```

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

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

Important features identification

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

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

  formula_single <- reformulate(fname, "class")

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

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

```
## # A tibble: 63 x 2
##    feat_name                                  p_value
##    <chr>                                        <dbl>
##  1 RuleAbstractNouns                         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 RuleDoubleAdpos.max_allowable_distance    2.74e- 4
##  8 RuleDoubleAdpos.max_allowable_distance.v  5.26e- 6
##  9 RuleInfVerbDistance                       5.24e-15
## 10 RuleInfVerbDistance.max_distance          5.48e- 2
## # i 53 more rows
```

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

Correlations

See Levshina (2015: 353–54).

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

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

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

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

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

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

what unites the low-communality variables we threw out:

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

High correlations

```
.hcorrcutoff <- 0.9

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

```
## # A tibble: 22 x 4
##    feat1                        feat2                          cor abs_cor
##    <chr>                        <chr>                        <dbl>   <dbl>
##  1 RuleLongSentences.max_length ari                          0.943   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

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

```
## Compare row 7  and column  34 with corr  0.943
##   Means:  0.399 vs 0.208 so flagging column 7
## Compare row 34  and column  40 with corr  0.978
##   Means:  0.382 vs 0.2 so flagging column 34
## Compare row 40  and column  48 with corr  0.987
##   Means:  0.368 vs 0.193 so flagging column 40
## Compare row 48  and column  38 with corr  0.948
##   Means:  0.348 vs 0.186 so flagging column 48
## Compare row 35  and column  36 with corr  0.96
##   Means:  0.26 vs 0.182 so flagging column 35
## Compare row 50  and column  41 with corr  0.958
##   Means:  0.185 vs 0.179 so flagging column 50
## Compare row 42  and column  45 with corr  0.964
##   Means:  0.174 vs 0.179 so flagging column 45
## All correlations <= 0.9
```

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

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

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

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

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

Low correlations

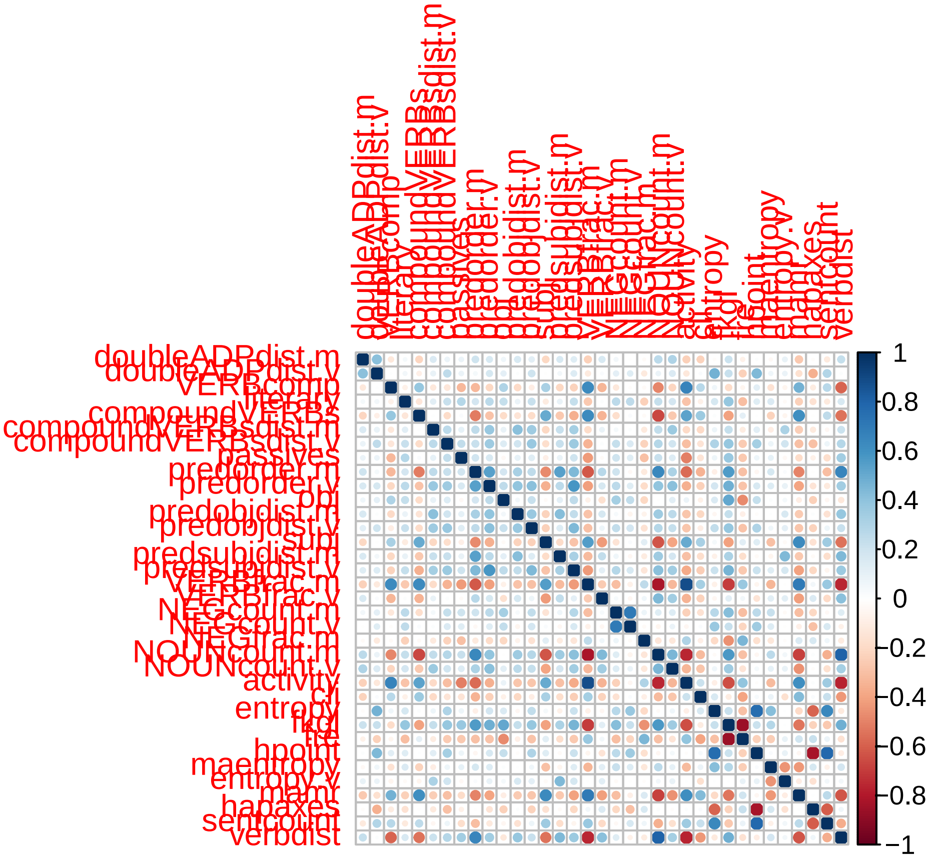
```r
# 0.35 instead of 0.3 otherwise the FA bootstrapping would freeze
.lcorrcutoff <- 0.35

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
```

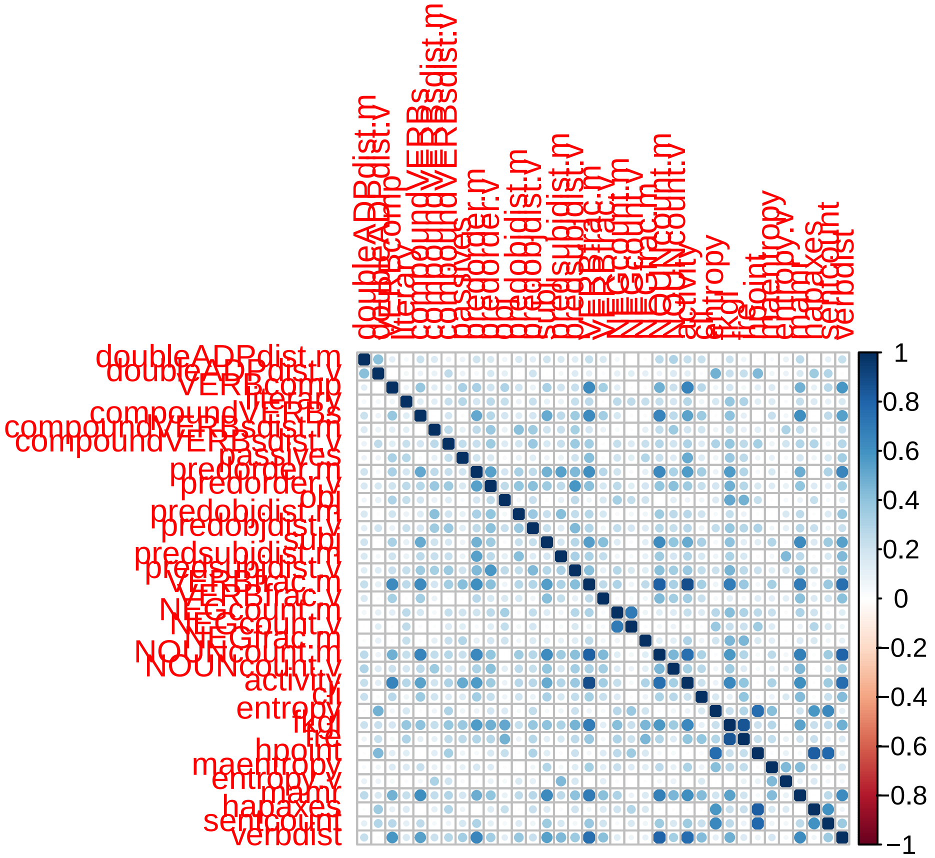
```r
data_pure <- data_pureish_striphigh %>%
  select(!any_of(low_correlating_features))

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

colnames(data_pure) <- cnames
```

Visualisation

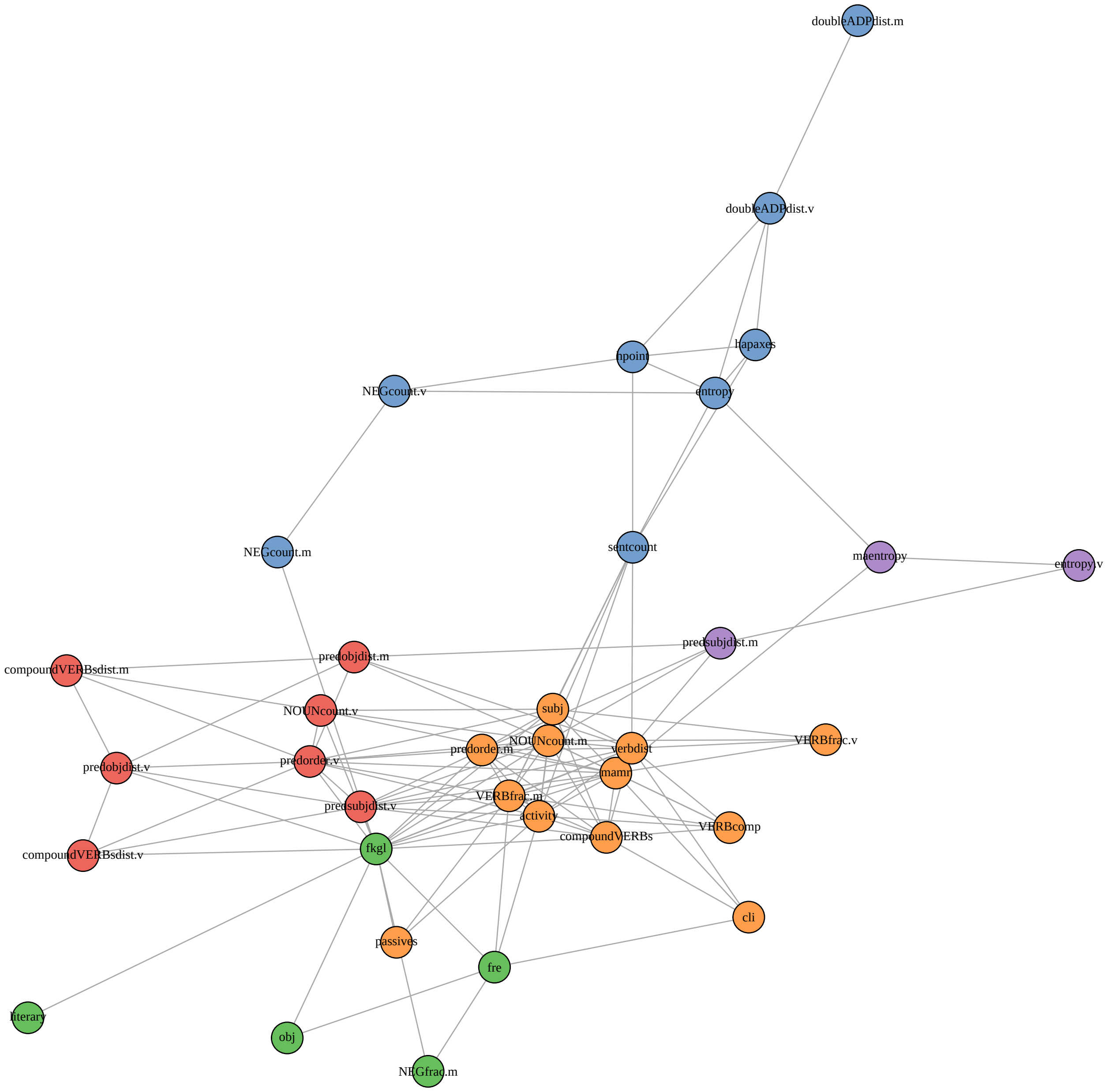
```r
corrplot(cor(data_pure))
```



```
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
```

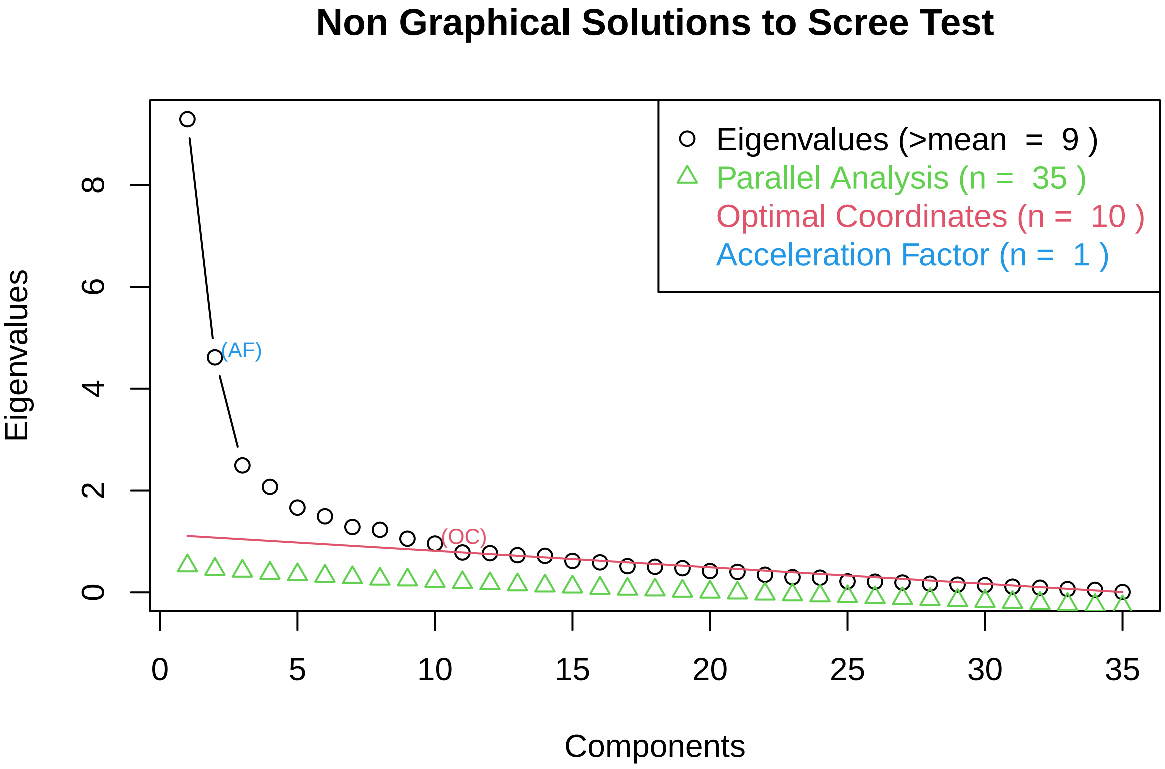
```
##          Beta-hat      kappa p-val
## Skewness 1171.473 147019.868     0
## Kurtosis 2988.432    456.547     0
```

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

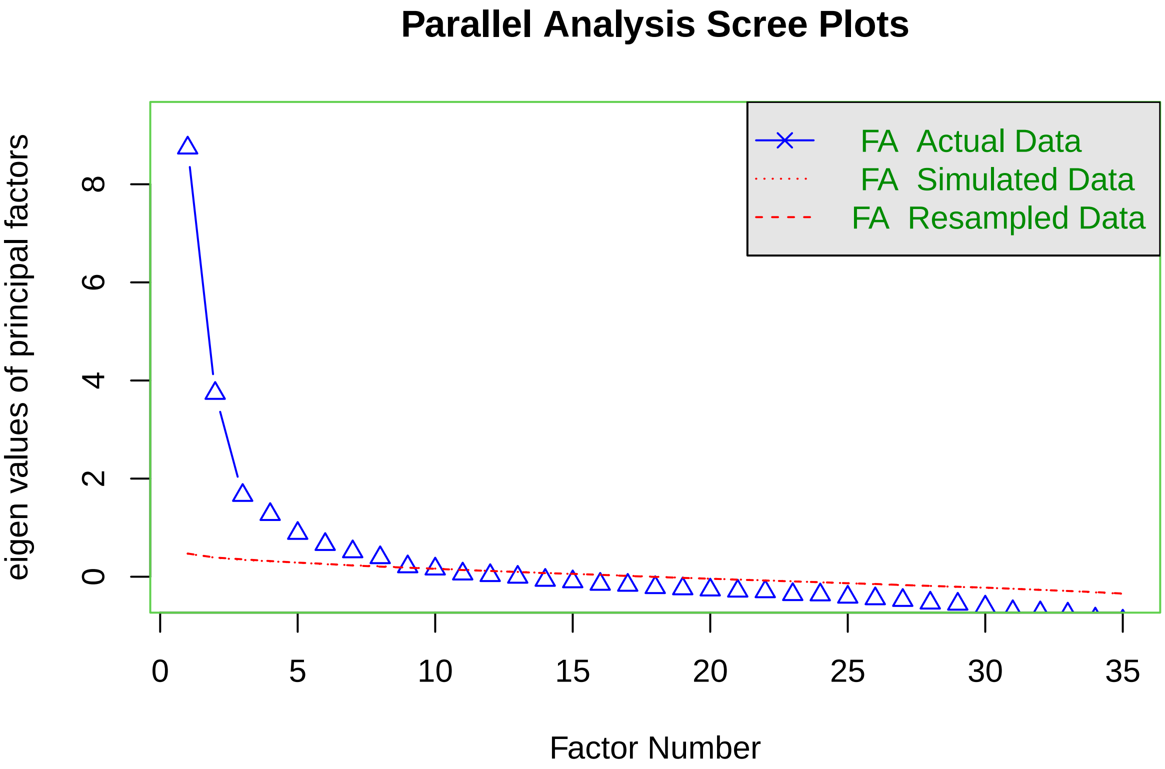
No. of factors

```
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 = 8,
  fm = "pa",
  rotate = "promax",
  oblique.scores = TRUE,
  scores = "tenBerge",
  n.iter = 100
)
```

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

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

```
fa_1
```

```
## Factor Analysis with confidence intervals using method = fa(r = data_scaled, nfactors = 8, n.iter =
##     scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Factor Analysis using method =  pa
## Call: fa(r = data_scaled, nfactors = 8, n.iter = 100, rotate = "promax",
##     scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##                       PA1   PA2   PA7   PA4   PA6   PA5   PA3   PA8   h2    u2
## doubleADPdist.m     -0.24  0.06  0.11  0.07  0.00 -0.10  0.04 -0.14 0.14 0.864
## doubleADPdist.v     -0.10  0.50  0.06  0.03  0.01 -0.08 -0.04  0.01 0.27 0.728
## VERBcomp             0.63  0.02  0.06  0.51  0.32 -0.11 -0.02  0.07 0.60 0.405
## literary             0.02 -0.03  0.12  0.17 -0.27  0.15 -0.08 -0.04 0.23 0.765
## compoundVERBs        1.04 -0.13  0.40 -0.26 -0.31  0.00 -0.01  0.13 0.73 0.273
## compoundVERBsdist.m  0.24 -0.04  0.81 -0.10 -0.10 -0.07  0.12 -0.04 0.47 0.528
## compoundVERBsdist.v -0.08  0.25  0.26  0.00 -0.19  0.05  0.09 -0.01 0.33 0.673
## passives             0.04 -0.07  0.04 -0.21 -0.84  0.07  0.00 -0.11 0.56 0.440
## predorder.m         -0.63 -0.12  0.09  0.20  0.05  0.01  0.13  0.04 0.59 0.414
## predorder.v         -0.07 -0.02  0.60  0.13  0.01  0.08 -0.03 -0.02 0.52 0.476
## obj                  0.14 -0.06 -0.02  0.93  0.19  0.15  0.08 -0.07 0.70 0.297
## predobjdist.m       -0.04 -0.13  0.60 -0.11  0.01 -0.04  0.15  0.08 0.38 0.621
## predobjdist.v        0.03  0.14  0.50  0.05 -0.03  0.08  0.01  0.04 0.37 0.634
## subj                 0.60  0.13 -0.19 -0.03 -0.12  0.05  0.22  0.08 0.54 0.457
## predsubjdist.m      -0.37 -0.07  0.20  0.06  0.15  0.03  0.46  0.22 0.51 0.494
## predsubjdist.v      -0.18  0.08  0.39  0.13 -0.01  0.12  0.06 -0.03 0.45 0.554
## VERBfrac.m           0.87 -0.05  0.20  0.02  0.36 -0.03  0.03  0.05 0.90 0.098
## VERBfrac.v          -0.54 -0.03  0.14 -0.21  0.24  0.00 -0.06  0.03 0.34 0.662
## NEGcount.m          -0.06 -0.07 -0.05  0.19  0.03  0.93 -0.02 -0.04 0.92 0.080
## NEGcount.v           0.19  0.10  0.03  0.07 -0.04  0.71 -0.05  0.00 0.59 0.409
## NEGfrac.m           -0.08 -0.06 -0.10 -0.20  0.41  0.26  0.09 -0.11 0.37 0.630
## NOUNcount.m         -0.91  0.02  0.00 -0.02 -0.02 -0.11 -0.03  0.04 0.81 0.187
## NOUNcount.v         -0.17 -0.04  0.48  0.03  0.01 -0.04 -0.12 -0.13 0.37 0.629
## activity             0.79 -0.03  0.12  0.23  0.49  0.02  0.05 -0.09 0.92 0.076
## cli                  0.33  0.04 -0.06 -0.05  0.13 -0.05 -0.14  0.75 0.72 0.283
## entropy              0.02  0.81  0.14 -0.15  0.08  0.10 -0.33  0.17 0.87 0.133
## fkgl                -0.40 -0.02 -0.03  0.57 -0.24  0.05 -0.03  0.17 0.96 0.037
## fre                  0.10  0.00  0.07 -0.55  0.14 -0.02  0.06 -0.59 0.97 0.032
## hpoint               0.00  0.97 -0.08  0.01 -0.02  0.06  0.08 -0.04 0.94 0.060
## maentropy           -0.33  0.04  0.03 -0.10  0.09  0.10 -0.72  0.23 0.68 0.321
## entropy.v            0.01  0.11  0.27 -0.03  0.11  0.01  0.53 -0.03 0.39 0.607
## mamr                 0.74 -0.08 -0.11  0.00 -0.05 -0.04  0.24  0.15 0.74 0.265
## hapaxes              0.01 -0.80  0.18 -0.11  0.10 -0.04 -0.26  0.10 0.73 0.268
## sentcount            0.19  0.91 -0.05 -0.24  0.23 -0.03  0.04  0.08 0.85 0.152
## verbdist            -0.84 -0.05 -0.02 -0.20 -0.21 -0.04  0.10 -0.04 0.79 0.210
##                     com
## doubleADPdist.m     3.1
## doubleADPdist.v     1.2
## VERBcomp            2.6
## literary            3.1
## compoundVERBs       1.7
## compoundVERBsdist.m 1.3
## compoundVERBsdist.v 3.4
## passives            1.2
## predorder.m         1.4
## predorder.v         1.2
## obj                 1.2
## predobjdist.m       1.4
## predobjdist.v       1.3
## subj                1.7
## predsubjdist.m      3.2
## predsubjdist.v      2.1
## VERBfrac.m          1.5
## VERBfrac.v          1.9
## NEGcount.m          1.1
## NEGcount.v          1.2
## NEGfrac.m           2.9
## NOUNcount.m         1.0
## NOUNcount.v         1.6
## activity            2.0
## cli                 1.6
## entropy             1.6
## fkgl                2.5
## fre                 2.2
## hpoint              1.0
## maentropy           1.8
## entropy.v           1.7
## mamr                1.4
## hapaxes             1.4
## sentcount           1.4
## verbdist            1.3
##
##                        PA1  PA2  PA7  PA4  PA6  PA5  PA3  PA8
## SS loadings           6.83 3.42 2.39 2.16 1.98 1.70 1.49 1.27
## Proportion Var        0.20 0.10 0.07 0.06 0.06 0.05 0.04 0.04
## Cumulative Var        0.20 0.29 0.36 0.42 0.48 0.53 0.57 0.61
## Proportion Explained  0.32 0.16 0.11 0.10 0.09 0.08 0.07 0.06
## Cumulative Proportion 0.32 0.48 0.60 0.70 0.79 0.87 0.94 1.00
##
##  With factor correlations of
##       PA1   PA2   PA7   PA4   PA6   PA5   PA3   PA8
## PA1  1.00  0.01 -0.62 -0.28  0.39 -0.20  0.00 -0.03
## PA2  0.01  1.00  0.27  0.34 -0.24  0.28 -0.05  0.12
## PA7 -0.62  0.27  1.00  0.43 -0.36  0.22  0.10  0.03
## PA4 -0.28  0.34  0.43  1.00 -0.44  0.22 -0.10  0.20
## PA6  0.39 -0.24 -0.36 -0.44  1.00 -0.26  0.06 -0.33
## PA5 -0.20  0.28  0.22  0.22 -0.26  1.00 -0.11  0.05
## PA3  0.00 -0.05  0.10 -0.10  0.06 -0.11  1.00 -0.03
## PA8 -0.03  0.12  0.03  0.20 -0.33  0.05 -0.03  1.00
##
## Mean item complexity =  1.8
## Test of the hypothesis that 8 factors are sufficient.
##
## df null model =  595  with the objective function =  28.69 with Chi Square =  21213.79
## df of  the model are 343  and the objective function was  4.9
##
## 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  919.3  with prob <  3.5e-54
## The total n.obs was  753  with Likelihood Chi Square =  3597.89  with prob <  0
##
## Tucker Lewis Index of factoring reliability =  0.724
## RMSEA index =  0.112  and the 90 % confidence intervals are  0.109 0.116
## BIC =  1325.83
## Fit based upon off diagonal values = 0.99
##  Coefficients and bootstrapped confidence intervals 
##                       low   PA1 upper   low   PA2 upper   low   PA7 upper   low
## doubleADPdist.m     -0.84 -0.24  0.27 -0.08  0.06  0.22 -0.62  0.11  0.99 -0.21
## doubleADPdist.v     -0.32 -0.10  0.08 -0.11  0.50  1.26 -0.32  0.06  0.49 -0.17
## VERBcomp            -0.50  0.63  1.97 -0.05  0.02  0.10 -0.23  0.06  0.39 -2.85
## literary            -0.14  0.02  0.11 -0.15 -0.03  0.06 -0.35  0.12  0.61 -1.05
## compoundVERBs       -0.89  1.04  3.25 -0.27 -0.13  0.04 -0.23  0.40  0.96 -2.35
## compoundVERBsdist.m -0.16  0.24  0.65 -0.13 -0.04  0.07 -0.98  0.81  2.82 -0.77
## compoundVERBsdist.v -0.32 -0.08  0.11 -0.02  0.25  0.59 -0.10  0.26  0.68 -0.25
## passives            -0.24  0.04  0.17 -0.17 -0.07  0.03 -0.43  0.04  0.48 -2.63
## predorder.m         -1.99 -0.63  0.50 -0.27 -0.12  0.03 -0.36  0.09  0.65 -1.36
## predorder.v         -0.34 -0.07  0.10 -0.11 -0.02  0.09 -0.65  0.60  2.01 -1.36
## obj                 -0.13  0.14  0.44 -0.16 -0.06  0.01 -0.44 -0.02  0.52 -4.70
## predobjdist.m       -0.28 -0.04  0.20 -0.33 -0.13  0.05 -0.63  0.60  2.12 -0.68
## predobjdist.v       -0.16  0.03  0.18 -0.05  0.14  0.38 -0.47  0.50  1.57 -0.55
## subj                -0.62  0.60  2.14 -0.05  0.13  0.32 -0.50 -0.19  0.11 -0.38
## predsubjdist.m      -0.92 -0.37  0.15 -0.16 -0.07  0.02 -0.34  0.20  0.88 -1.62
## predsubjdist.v      -0.61 -0.18  0.14 -0.03  0.08  0.22 -0.51  0.39  1.42 -1.25
## VERBfrac.m          -0.80  0.87  2.88 -0.13 -0.05  0.03 -0.08  0.20  0.49 -0.54
## VERBfrac.v          -1.74 -0.54  0.47 -0.15 -0.03  0.07 -0.13  0.14  0.42 -1.13
## NEGcount.m          -0.14 -0.06  0.07 -0.17 -0.07  0.05 -0.33 -0.05  0.22 -1.73
## NEGcount.v          -0.22  0.19  0.71 -0.04  0.10  0.26 -0.19  0.03  0.22 -1.28
## NEGfrac.m           -0.15 -0.08  0.12 -0.21 -0.06  0.07 -0.26 -0.10  0.06 -1.39
## NOUNcount.m         -2.95 -0.91  0.82 -0.05  0.02  0.09 -0.24  0.00  0.31 -0.71
## NOUNcount.v         -0.76 -0.17  0.28 -0.17 -0.04  0.06 -0.50  0.48  1.63 -0.61
## activity            -0.65  0.79  2.51 -0.08 -0.03  0.02 -0.08  0.12  0.33 -1.41
## cli                 -0.50  0.33  1.49 -0.05  0.04  0.16 -0.58 -0.06  0.38 -1.16
## entropy             -0.09  0.02  0.08 -0.20  0.81  2.06 -0.03  0.14  0.25 -0.99
## fkgl                -1.36 -0.40  0.41 -0.09 -0.02  0.02 -0.22 -0.03  0.21 -3.25
## fre                 -0.10  0.10  0.22 -0.06  0.00  0.05 -0.23  0.07  0.34 -6.20
## hpoint              -0.08  0.00  0.12 -0.26  0.97  2.45 -0.25 -0.08  0.15 -0.42
## maentropy           -1.13 -0.33  0.27 -0.05  0.04  0.18 -0.35  0.03  0.32 -0.71
## entropy.v           -0.22  0.01  0.26 -0.02  0.11  0.28 -0.28  0.27  0.88 -0.59
## mamr                -0.79  0.74  2.66 -0.24 -0.08  0.05 -0.32 -0.11  0.09 -0.28
## hapaxes             -0.23  0.01  0.22 -2.01 -0.80  0.21 -0.10  0.18  0.39 -0.66
## sentcount           -0.24  0.19  0.79 -0.25  0.91  2.32 -0.25 -0.05  0.16 -2.05
## verbdist            -2.72 -0.84  0.72 -0.13 -0.05  0.03 -0.18 -0.02  0.18 -1.98
##                       PA4 upper   low   PA6 upper   low   PA5 upper   low   PA3
## doubleADPdist.m      0.07  0.38 -0.45  0.00  0.57 -0.82 -0.10  0.42 -0.38  0.04
## doubleADPdist.v      0.03  0.21 -0.26  0.01  0.35 -1.11 -0.08  0.71 -0.62 -0.04
## VERBcomp             0.51  4.76 -2.16  0.32  3.61 -0.79 -0.11  0.45 -0.16 -0.02
## literary             0.17  1.66 -1.72 -0.27  0.89 -1.11  0.15  1.92 -0.56 -0.08
## compoundVERBs       -0.26  1.42 -2.96 -0.31  1.80 -0.66  0.00  0.59 -0.24 -0.01
## compoundVERBsdist.m -0.10  0.48 -0.96 -0.10  0.59 -0.65 -0.07  0.36 -0.95  0.12
## compoundVERBsdist.v  0.00  0.28 -1.88 -0.19  1.15 -0.54  0.05  0.80 -0.72  0.09
## passives            -0.21  1.73 -6.31 -0.84  3.54 -0.73  0.07  1.23 -0.28  0.00
## predorder.m          0.20  2.18 -0.53  0.05  0.55 -0.54  0.01  0.48 -0.42  0.13
## predorder.v          0.13  1.96 -0.24  0.01  0.31 -0.25  0.08  0.58 -0.26 -0.03
## obj                  0.93  8.04 -1.48  0.19  2.48 -1.57  0.15  2.45 -0.52  0.08
## predobjdist.m       -0.11  0.36 -0.79  0.01  0.62 -1.02 -0.04  0.70 -0.52  0.15
## predobjdist.v        0.05  0.86 -0.71 -0.03  0.54 -0.42  0.08  0.70 -0.26  0.01
## subj                -0.03  0.24 -1.19 -0.12  0.68 -0.17  0.05  0.38 -0.93  0.22
## predsubjdist.m       0.06  2.11 -0.44  0.15  0.60 -1.03  0.03  0.87 -1.63  0.46
## predsubjdist.v       0.13  1.86 -0.19 -0.01  0.18 -0.60  0.12  1.12 -0.30  0.06
## VERBfrac.m           0.02  0.71 -1.77  0.36  3.15 -0.67 -0.03  0.48 -0.14  0.03
## VERBfrac.v          -0.21  0.51 -0.96  0.24  1.71 -0.80  0.00  0.62 -1.08 -0.06
## NEGcount.m           0.19  2.66 -0.76  0.03  0.53 -4.88  0.93  8.44 -0.47 -0.02
## NEGcount.v           0.07  1.74 -0.83 -0.04  0.51 -3.80  0.71  6.73 -0.77 -0.05
## NEGfrac.m           -0.20  0.79 -1.57  0.41  2.83 -1.00  0.26  1.86 -0.45  0.09
## NOUNcount.m         -0.02  0.61 -0.61 -0.02  0.35 -1.36 -0.11  0.80 -0.36 -0.03
## NOUNcount.v          0.03  0.67 -0.61  0.01  0.83 -0.24 -0.04  0.24 -0.95 -0.12
## activity             0.23  2.32 -2.84  0.49  4.91 -0.18  0.02  0.30 -0.30  0.05
## cli                 -0.05  1.40 -1.15  0.13  0.89 -1.93 -0.05  1.27 -1.99 -0.14
## entropy             -0.15  0.48 -0.18  0.08  0.45 -0.37  0.10  0.79 -3.01 -0.33
## fkgl                 0.57  5.43 -2.04 -0.24  1.14 -0.67  0.05  0.98 -0.23 -0.03
## fre                 -0.55  3.89 -1.58  0.14  2.58 -0.31 -0.02  0.46 -0.58  0.06
## hpoint               0.01  0.38 -0.16 -0.02  0.17 -0.52  0.06  0.86 -0.40  0.08
## maentropy           -0.10  0.38 -0.21  0.09  0.59 -0.18  0.10  0.58 -6.55 -0.72
## entropy.v           -0.03  0.61 -0.29  0.11  0.51 -0.32  0.01  0.21 -2.50  0.53
## mamr                 0.00  0.35 -1.02 -0.05  0.66 -1.02 -0.04  0.70 -1.03  0.24
## hapaxes             -0.11  0.35 -0.23  0.10  0.51 -0.99 -0.04  0.67 -2.35 -0.26
## sentcount           -0.24  1.18 -0.84  0.23  1.59 -0.93 -0.03  0.65 -0.15  0.04
## verbdist            -0.20  1.21 -2.43 -0.21  1.44 -0.81 -0.04  0.57 -0.33  0.10
##                     upper   low   PA8 upper
## doubleADPdist.m      0.60 -1.28 -0.14  0.63
## doubleADPdist.v      0.41 -1.36  0.01  1.76
## VERBcomp             0.16 -2.28  0.07  3.32
## literary             0.37 -1.09 -0.04  1.49
## compoundVERBs        0.28 -2.24  0.13  3.44
## compoundVERBsdist.m  1.63 -1.88 -0.04  1.50
## compoundVERBsdist.v  1.20 -0.74 -0.01  0.82
## passives             0.48 -1.76 -0.11  2.12
## predorder.m          0.89 -0.83  0.04  0.68
## predorder.v          0.27 -0.51 -0.02  0.44
## obj                  0.97 -1.79 -0.07  2.21
## predobjdist.m        1.10 -1.19  0.08  1.18
## predobjdist.v        0.40 -0.35  0.04  0.46
## subj                 1.80 -0.64  0.08  1.05
## predsubjdist.m       3.24 -1.59  0.22  1.74
## predsubjdist.v       0.62 -1.00 -0.03  0.65
## VERBfrac.m           0.25 -0.31  0.05  0.44
## VERBfrac.v           0.64 -2.75  0.03  2.07
## NEGcount.m           0.21 -4.51 -0.04  3.23
## NEGcount.v           0.40 -2.54  0.00  1.85
## NEGfrac.m            0.71 -6.42 -0.11  4.41
## NOUNcount.m          0.18 -0.76  0.04  0.59
## NOUNcount.v          0.54 -2.10 -0.13  1.53
## activity             0.61 -1.45 -0.09  0.94
## cli                  1.03 -4.30  0.75  7.45
## entropy              1.58 -1.42  0.17  2.60
## fkgl                 0.11 -3.59  0.17  5.27
## fre                  1.11 -9.81 -0.59  6.30
## hpoint               0.76 -0.53 -0.04  0.47
## maentropy            3.39 -2.39  0.23  4.19
## entropy.v            4.68 -4.05 -0.03  2.89
## mamr                 1.97 -1.60  0.15  2.36
## hapaxes              1.19 -0.46  0.10  0.90
## sentcount            0.22 -0.43  0.08  0.53
## verbdist             0.69 -1.76 -0.04  1.15
##
##  Interfactor correlations and bootstrapped confidence intervals
##         lower estimate upper
## PA1-PA2 -0.30   0.0086  0.37
## PA1-PA7 -0.74  -0.6249  0.49
## PA1-PA4 -1.00  -0.2794  0.25
## PA1-PA6 -0.85   0.3930  0.54
## PA1-PA5 -0.64  -0.1972  0.38
## PA1-PA3 -0.52   0.0028  0.36
## PA1-PA8 -0.37  -0.0280  0.25
## PA2-PA7 -0.32   0.2716  0.61
## PA2-PA4 -0.14   0.3420  0.51
## PA2-PA6 -0.21  -0.2390  0.61
## PA2-PA5 -0.23   0.2815  0.44
## PA2-PA3 -0.26  -0.0492  0.34
## PA2-PA8 -0.24   0.1202  0.24
## PA7-PA4 -0.53   0.4325  0.84
## PA7-PA6 -0.68  -0.3594  0.81
## PA7-PA5 -0.44   0.2202  0.52
## PA7-PA3 -0.41   0.0986  0.40
## PA7-PA8 -0.35   0.0276  0.30
## PA4-PA6 -0.50  -0.4380  0.78
## PA4-PA5 -0.37   0.2155  0.57
## PA4-PA3 -0.35  -0.1018  0.47
## PA4-PA8 -0.37   0.2047  0.37
## PA6-PA5 -0.35  -0.2569  0.42
## PA6-PA3 -0.36   0.0611  0.39
## PA6-PA8 -0.35  -0.3300  0.27
## PA5-PA3 -0.34  -0.1103  0.36
## PA5-PA8 -0.34   0.0506  0.25
## PA3-PA8 -0.32  -0.0318  0.29
```

Healthiness diagnostics

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

```
## # A tibble: 35 x 2
##    feat                maxload
##    <chr>                 <dbl>
##  1 doubleADPdist.m       0.236
##  2 compoundVERBsdist.v   0.258
##  3 literary              0.271
##  4 predsubjdist.v        0.393
##  5 NEGfrac.m             0.412
##  6 predsubjdist.m        0.457
##  7 NOUNcount.v           0.485
##  8 doubleADPdist.v       0.497
##  9 predobjdist.v         0.498
## 10 entropy.v             0.528
## # i 25 more rows
```

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

```
##     doubleADPdist.m            literary     doubleADPdist.v compoundVERBsdist.v
##           0.1357450           0.2347000           0.2722296           0.3272989
##          VERBfrac.v       predobjdist.v           NEGfrac.m         NOUNcount.v
##           0.3380620           0.3659035           0.3697819           0.3712940
##       predobjdist.m           entropy.v      predsubjdist.v compoundVERBsdist.m
##           0.3789752           0.3932817           0.4457381           0.4723022
##      predsubjdist.m         predorder.v                subj            passives
##           0.5058605           0.5239791           0.5434694           0.5595964
##         predorder.m          NEGcount.v            VERBcomp           maentropy
##           0.5856440           0.5914438           0.5951878           0.6791988
##                 obj                 cli       compoundVERBs             hapaxes
##           0.7034850           0.7165486           0.7272900           0.7318621
##                mamr            verbdist         NOUNcount.m           sentcount
##           0.7351373           0.7896616           0.8131161           0.8482973
##             entropy          VERBfrac.m          NEGcount.m            activity
##           0.8666303           0.9020945           0.9196779           0.9242143
##              hpoint                fkgl                 fre
##           0.9404694           0.9632766           0.9679727
```

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

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

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

```
##              hpoint         NOUNcount.m          NEGcount.m         predorder.v
##            1.038219            1.038952            1.118903            1.177659
##     doubleADPdist.v            passives                 obj          NEGcount.v
##            1.191882            1.200888            1.224012            1.228227
##       predobjdist.v            verbdist compoundVERBsdist.m       predobjdist.m
##            1.263473            1.294983            1.320354            1.356689
##                mamr           sentcount         predorder.m             hapaxes
##            1.396840            1.416152            1.441863            1.442406
##          VERBfrac.m         NOUNcount.v                 cli             entropy
##            1.478726            1.569636            1.578381            1.643849
##       compoundVERBs           entropy.v                subj           maentropy
##            1.709315            1.725227            1.743929            1.764165
##          VERBfrac.v            activity      predsubjdist.v                 fre
##            1.919053            1.990248            2.102798            2.232114
##                fkgl            VERBcomp           NEGfrac.m     doubleADPdist.m
##            2.462100            2.575866            2.897845            3.067314
##            literary      predsubjdist.m compoundVERBsdist.v
##            3.077866            3.236255            3.409573
```

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

```
## [1] "doubleADPdist.m"     "VERBcomp"            "literary"
## [4] "compoundVERBsdist.v" "predsubjdist.m"      "predsubjdist.v"
## [7] "NEGfrac.m"           "fkgl"                "fre"
```

Feature engineering

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

det(cor(data_engineered_1))
```

```
## [1] 1.324081e-07
```

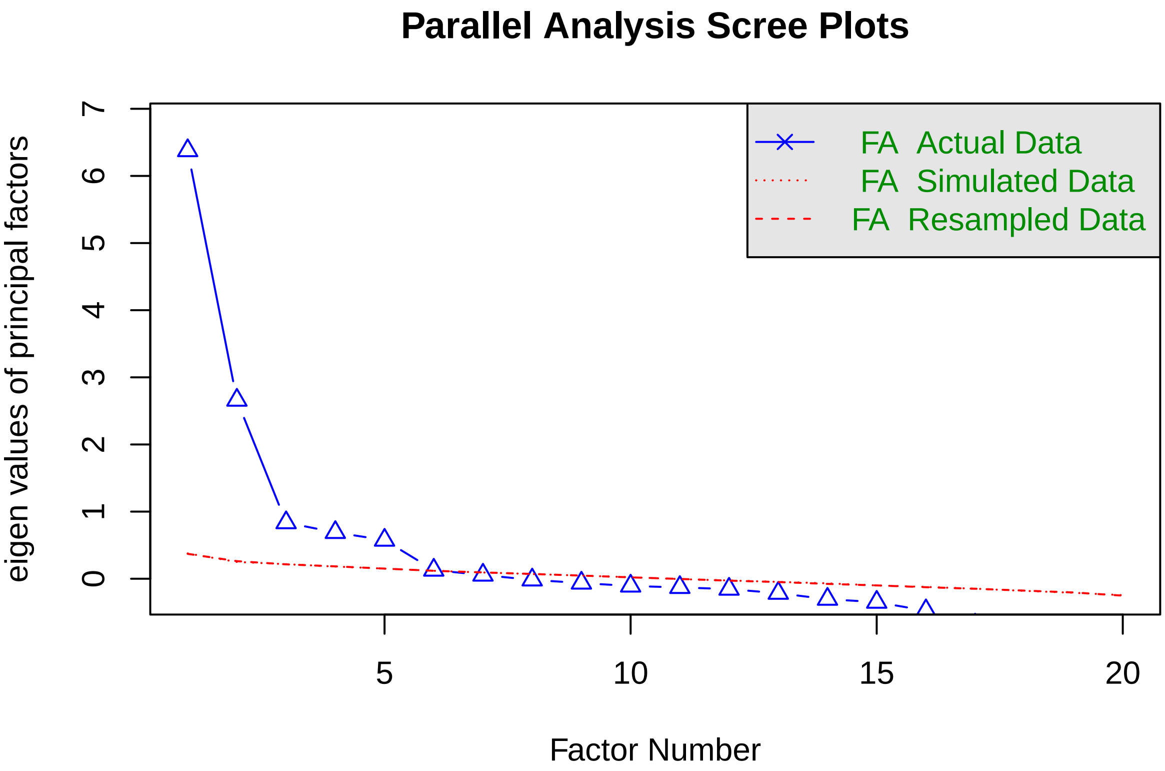
```r
KMO(data_engineered_1)
```

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

second FA

No. of vectors

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



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

Model

```
set.seed(42)

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

```
## Factor Analysis with confidence intervals using method = fa(r = data_engineered_1, nfactors = 5, n.i
##     scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Factor Analysis using method =  pa
## Call: fa(r = data_engineered_1, nfactors = 5, n.iter = 100, rotate = "promax",
##     scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##                 PA1   PA2   PA4   PA3   PA5   h2    u2 com
## VERBcomp       0.23  0.05  0.61  0.05 -0.04 0.56 0.436 1.3
## compoundVERBs  0.75  0.00 -0.12  0.09 -0.18 0.55 0.453 1.2
## passives       0.03  0.01 -0.60  0.23 -0.12 0.35 0.653 1.4
## predorder.m   -0.85 -0.03  0.02  0.00 -0.16 0.69 0.314 1.1
## predorder.v   -0.54  0.10  0.05  0.16 -0.02 0.35 0.651 1.3
## obj           -0.31  0.00  0.44  0.41 -0.06 0.46 0.543 2.8
## subj            0.61  0.14 -0.07  0.05 -0.29 0.52 0.482 1.6
## predsubjdist.m -0.54  0.02 -0.02 -0.04 -0.28 0.30 0.696 1.5
## VERBfrac.m      0.64 -0.04  0.42 -0.08 -0.10 0.88 0.118 1.8
## NEGcount.m      0.03 -0.10 -0.16  0.89  0.14 0.76 0.242 1.1
## NEGcount.v      0.25  0.05 -0.18  0.79  0.11 0.62 0.379 1.4
## NOUNcount.m    -0.82  0.04 -0.16 -0.17  0.10 0.81 0.194 1.2
## activity        0.49 -0.05  0.61 -0.02 -0.07 0.89 0.110 2.0
## entropy         0.03  0.76  0.03  0.10  0.46 0.86 0.145 1.7
## hpoint         -0.10  0.98 -0.03  0.03 -0.03 0.96 0.038 1.0
## maentropy      -0.09 -0.02  0.06  0.13  0.71 0.53 0.465 1.1
## mamr            0.65 -0.03  0.03 -0.03 -0.40 0.72 0.282 1.7
## hapaxes         0.14 -0.83  0.07 -0.04  0.25 0.75 0.255 1.3
## sentcount       0.22  0.87  0.10 -0.22  0.03 0.82 0.185 1.3
## verbdist       -0.69 -0.01 -0.39 -0.14 -0.06 0.79 0.210 1.7
##
##                        PA1  PA2  PA4  PA3  PA5
## SS loadings           5.08 3.00 2.04 1.72 1.31
## Proportion Var        0.25 0.15 0.10 0.09 0.07
## Cumulative Var        0.25 0.40 0.51 0.59 0.66
## Proportion Explained  0.39 0.23 0.16 0.13 0.10
## Cumulative Proportion 0.39 0.61 0.77 0.90 1.00
##
##  With factor correlations of
##       PA1   PA2   PA4   PA3   PA5
## PA1  1.00  0.11  0.39 -0.22 -0.21
## PA2  0.11  1.00  0.14  0.37 -0.01
## PA4  0.39  0.14  1.00  0.09 -0.32
## PA3 -0.22  0.37  0.09  1.00  0.00
## PA5 -0.21 -0.01 -0.32  0.00  1.00
##
## Mean item complexity =  1.5
## Test of the hypothesis that 5 factors are sufficient.
##
## df null model =  190  with the objective function =  15.84 with Chi Square =  11790.93
## df of  the model are 100  and the objective function was  1.88
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.05
##
## The harmonic n.obs is  753 with the empirical chi square  334.48  with prob <  4.8e-27
## The total n.obs was  753  with Likelihood Chi Square =  1390.4  with prob <  3.9e-226
##
## Tucker Lewis Index of factoring reliability =  0.788
## RMSEA index =  0.131  and the 90 % confidence intervals are  0.125 0.137
## BIC =  728
## Fit based upon off diagonal values = 0.99
## Measures of factor score adequacy
##                                                  PA1  PA2  PA4  PA3 PA5
## Correlation of (regression) scores with factors   0.97 0.99 0.93 0.93 0.9
## Multiple R square of scores with factors          0.94 0.98 0.87 0.86 0.8
## Minimum correlation of possible factor scores     0.87 0.96 0.74 0.72 0.6
##
##  Coefficients and bootstrapped confidence intervals
##                  low   PA1 upper   low   PA2 upper   low   PA4 upper   low
## VERBcomp        0.14  0.23  0.33 -0.01  0.05  0.11  0.47  0.61  0.75  0.00
## compoundVERBs   0.56  0.75  0.93 -0.07  0.00  0.05 -0.21 -0.12 -0.04  0.02
## passives       -0.07  0.03  0.15 -0.06  0.01  0.08 -0.75 -0.60 -0.49  0.14
## predorder.m    -0.95 -0.85 -0.70 -0.08 -0.03  0.01 -0.10  0.02  0.12 -0.06
## predorder.v    -0.63 -0.54 -0.41  0.03  0.10  0.16 -0.05  0.05  0.14  0.08
## obj            -0.39 -0.31 -0.19 -0.05  0.00  0.06  0.33  0.44  0.56  0.33
## subj            0.49  0.61  0.72  0.07  0.14  0.20 -0.15 -0.07  0.02 -0.01
## predsubjdist.m -0.66 -0.54 -0.37 -0.03  0.02  0.06 -0.17 -0.02  0.10 -0.11
## VERBfrac.m      0.51  0.64  0.78 -0.07 -0.04  0.00  0.31  0.42  0.53 -0.13
## NEGcount.m     -0.06  0.03  0.08 -0.13 -0.10 -0.05 -0.24 -0.16 -0.08  0.82
## NEGcount.v      0.16  0.25  0.32  0.00  0.05  0.11 -0.26 -0.18 -0.09  0.73
## NOUNcount.m    -0.98 -0.82 -0.64  0.00  0.04  0.08 -0.24 -0.16 -0.08 -0.23
## activity        0.39  0.49  0.59 -0.08 -0.05 -0.01  0.49  0.61  0.74 -0.07
## entropy        -0.03  0.03  0.08  0.71  0.76  0.81 -0.04  0.03  0.10  0.06
## hpoint         -0.14 -0.10 -0.06  0.95  0.98  0.99 -0.06 -0.03  0.01  0.01
## maentropy      -0.18 -0.09  0.02 -0.07 -0.02  0.02 -0.02  0.06  0.14  0.07
## mamr            0.50  0.65  0.80 -0.07 -0.03  0.01 -0.03  0.03  0.10 -0.08
## hapaxes         0.08  0.14  0.19 -0.86 -0.83 -0.80  0.00  0.07  0.14 -0.09
## sentcount       0.16  0.22  0.28  0.83  0.87  0.91  0.05  0.10  0.15 -0.29
## verbdist       -0.79 -0.69 -0.57 -0.04 -0.01  0.02 -0.53 -0.39 -0.27 -0.24
##                  PA3 upper   low   PA5 upper
## VERBcomp        0.05  0.12 -0.12 -0.04  0.03
## compoundVERBs   0.09  0.18 -0.36 -0.18 -0.02
## passives        0.23  0.35 -0.27 -0.12  0.01
## predorder.m     0.00  0.09 -0.38 -0.16  0.09
## predorder.v     0.16  0.25 -0.09 -0.02  0.08
## obj             0.41  0.52 -0.13 -0.06  0.02
## subj            0.05  0.12 -0.47 -0.29 -0.14
## predsubjdist.m -0.04  0.04 -0.52 -0.28 -0.06
## VERBfrac.m     -0.08 -0.01 -0.20 -0.10 -0.02
## NEGcount.m      0.89  0.97  0.05  0.14  0.25
## NEGcount.v      0.79  0.86  0.03  0.11  0.21
## NOUNcount.m    -0.17 -0.12  0.01  0.10  0.20
## activity       -0.02  0.02 -0.12 -0.07 -0.03
## entropy         0.10  0.16  0.38  0.46  0.60
## hpoint          0.03  0.06 -0.07 -0.03  0.02
## maentropy       0.13  0.20  0.64  0.71  0.93
## mamr           -0.03  0.01 -0.64 -0.40 -0.20
## hapaxes        -0.04  0.01  0.19  0.25  0.33
## sentcount      -0.22 -0.17 -0.02  0.03  0.09
## verbdist       -0.14 -0.05 -0.14 -0.06  0.02
##
##  Interfactor correlations and bootstrapped confidence intervals
##         lower estimate upper
## PA1-PA2 -0.24   0.1089  0.31
## PA1-PA4 -0.78   0.3894  0.60
## PA1-PA3 -0.72  -0.2244  0.65
## PA1-PA5 -0.38  -0.2113  0.29
## PA2-PA4 -0.16   0.1385  0.52
## PA2-PA3 -0.20   0.3719  0.59
## PA2-PA5 -0.17  -0.0056  0.37
## PA4-PA3 -0.35   0.0862  0.41
## PA4-PA5 -0.37  -0.3247  0.44
## PA3-PA5 -0.23  -0.0043  0.25
```

Healthiness diagnostics

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

```
## # A tibble: 20 x 2
##    feat           maxload
##    <chr>            <dbl>
##  1 obj              0.444
##  2 predsubjdist.m   0.543
##  3 predorder.v      0.544
##  4 passives         0.604
##  5 VERBcomp         0.607
##  6 subj             0.611
##  7 activity         0.614
##  8 VERBfrac.m       0.644
##  9 mamr             0.650
## 10 verbdist         0.687
## 11 maentropy        0.713
## 12 compoundVERBs    0.745
## 13 entropy          0.764
## 14 NEGcount.v       0.793
## 15 NOUNcount.m      0.817
## 16 hapaxes          0.830
## 17 predorder.m      0.849
## 18 sentcount        0.871
## 19 NEGcount.m       0.889
## 20 hpoint           0.976
```

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

```
## predsubjdist.m       passives    predorder.v            obj           subj
##      0.3037802      0.3466316      0.3493952      0.4567127      0.5175456
##      maentropy  compoundVERBs       VERBcomp     NEGcount.v    predorder.m
##      0.5346480      0.5468982      0.5640821      0.6207903      0.6856355
##           mamr        hapaxes     NEGcount.m       verbdist    NOUNcount.m
##      0.7184074      0.7453044      0.7583936      0.7896065      0.8064764
##      sentcount        entropy     VERBfrac.m       activity         hpoint
##      0.8153085      0.8553720      0.8820377      0.8903719      0.9620798
```

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

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

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

```
##         hpoint    predorder.m      maentropy     NEGcount.m  compoundVERBs
##       1.026762       1.070995       1.111323       1.143923       1.200722
##    NOUNcount.m    predorder.v        hapaxes      sentcount       VERBcomp
##       1.203156       1.262321       1.268222       1.292409       1.310246
##     NEGcount.v       passives predsubjdist.m           subj           mamr
##       1.371912       1.387429       1.517215       1.593142       1.667229
##        entropy       verbdist     VERBfrac.m       activity            obj
##       1.687311       1.701920       1.820353       1.951844       2.820171
```

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

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

Feature engineering

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

det(cor(data_engineered_2))
```

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

```
KMO(data_engineered_2)
```

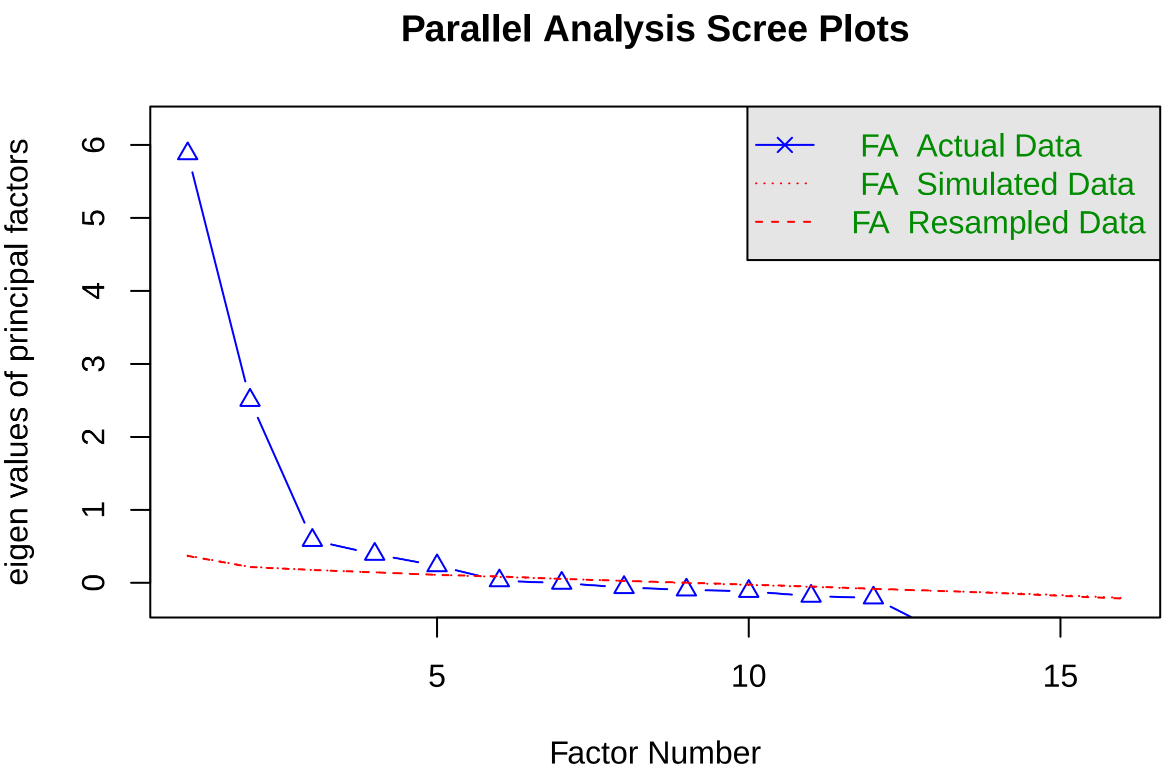
```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = data_engineered_2)
## Overall MSA =  0.84
## MSA for each item =
##      VERBcomp compoundVERBs   predorder.m          subj    VERBfrac.m
##          0.84          0.94          0.94          0.94          0.86
##    NEGcount.m    NEGcount.v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

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

Final FA

No. of vectors

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



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

Model

```r
set.seed(42)

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

```
## Factor Analysis with confidence intervals using method = fa(r = data_engineered_2, nfactors = 5, n.i
##     scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Factor Analysis using method =  pa
## Call: fa(r = data_engineered_2, nfactors = 5, n.iter = 100, rotate = "promax",
##     scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##                   PA1   PA2   PA5   PA3   PA4   h2    u2 com
## VERBcomp         0.15  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       0.02  0.15  0.28  0.04  0.09  0.14  0.44  0.60  0.81 -0.05  0.01
## compoundVERBs  0.71  0.79  0.86 -0.11 -0.06 -0.01 -0.18 -0.08  0.04 -0.04  0.02
## predorder.m   -0.89 -0.75 -0.66 -0.02  0.02  0.07 -0.10  0.02  0.12 -0.05  0.03
## subj           0.66  0.75  0.80  0.07  0.11  0.16 -0.26 -0.16 -0.03 -0.04  0.00
## VERBfrac.m     0.54  0.60  0.70 -0.09 -0.06 -0.03  0.32  0.44  0.55 -0.10 -0.06
## NEGcount.m    -0.15 -0.11 -0.06 -0.08 -0.05 -0.02 -0.01  0.04  0.08  0.84  0.91
## NEGcount.v     0.11  0.17  0.22  0.04  0.07  0.12 -0.09 -0.03  0.03  0.70  0.80
## NOUNcount.m   -0.98 -0.88 -0.80  0.04  0.07  0.11 -0.18 -0.09 -0.02 -0.16 -0.10
## activity       0.30  0.39  0.52 -0.06 -0.03 -0.01  0.46  0.65  0.85 -0.03  0.01
## entropy        0.05  0.10  0.13  0.67  0.71  0.75 -0.10 -0.06 -0.01 -0.02  0.01
## hpoint        -0.17 -0.13 -0.09  0.96  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 -0.07 -0.10 -0.03  0.03 -0.02  0.01
## mamr           0.65  0.74  0.82 -0.08 -0.04  0.01 -0.14 -0.02  0.12 -0.10 -0.05
## hapaxes        0.12  0.18  0.23 -0.91 -0.88 -0.86 -0.07 -0.01  0.05 -0.13 -0.08
## sentcount      0.15  0.21  0.30  0.77  0.80  0.83  0.02  0.09  0.17 -0.19 -0.15
## verbdist      -0.77 -0.69 -0.61 -0.03  0.00  0.03 -0.44 -0.29 -0.18 -0.13 -0.07
##               upper   low   PA4 upper
## VERBcomp       0.07 -0.07 -0.01  0.04
## compoundVERBs  0.08 -0.06  0.00  0.05
## predorder.m    0.12 -0.17 -0.12 -0.06
## subj           0.06 -0.21 -0.14 -0.08
## VERBfrac.m    -0.03 -0.06 -0.03  0.00
## NEGcount.m     1.00 -0.04  0.00  0.03
## NEGcount.v     0.87 -0.02  0.02  0.07
## NOUNcount.m   -0.05 -0.06 -0.02  0.02
## activity       0.04 -0.10 -0.06 -0.03
## entropy        0.05  0.50  0.55  0.60
## hpoint         0.08 -0.08 -0.05 -0.02
## maentropy      0.05  0.72  0.77  0.84
## mamr           0.00 -0.32 -0.26 -0.20
## hapaxes       -0.04  0.24  0.29  0.32
## 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  0.109    0.184  0.26
## PA1-PA5 -0.638    0.608  0.96
## PA1-PA3 -0.635   -0.165  0.86
## PA1-PA4 -0.603   -0.259  0.30
## PA2-PA5 -0.003    0.071  0.41
## PA2-PA3 -0.057    0.289  0.39
## PA2-PA4 -0.014    0.162  0.29
## PA5-PA3 -0.470   -0.173  0.28
## PA5-PA4 -0.389   -0.152  0.49
## PA3-PA4 -0.405    0.282  0.45
```

Healthiness diagnostics

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

```
## # A tibble: 16 x 2
##    feat          maxload
##    <chr>           <dbl>
##  1 VERBcomp        0.599
##  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
```

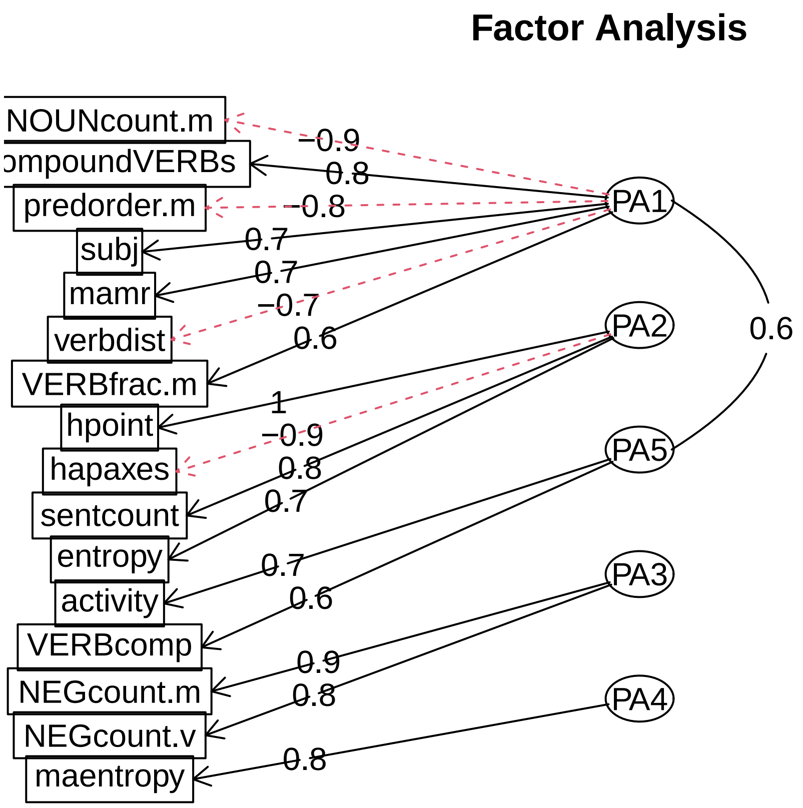
```
fa_res$complexity[fa_res$complexity > 2] %>% names()
```

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

Loadings

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

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



```
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
```

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

```r
  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

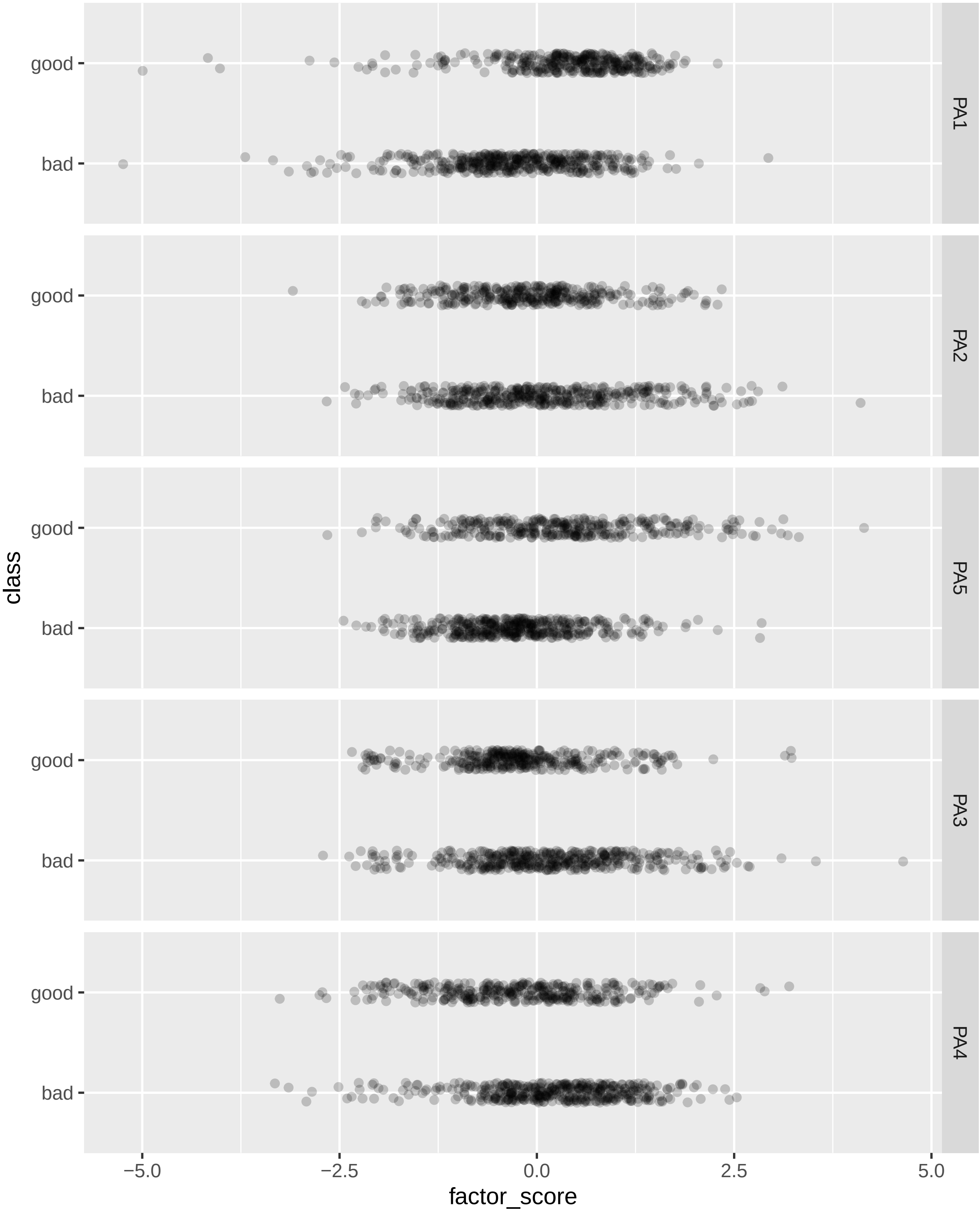
```
fa_res$uniquenesses %>% round(3)
```

```
##      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

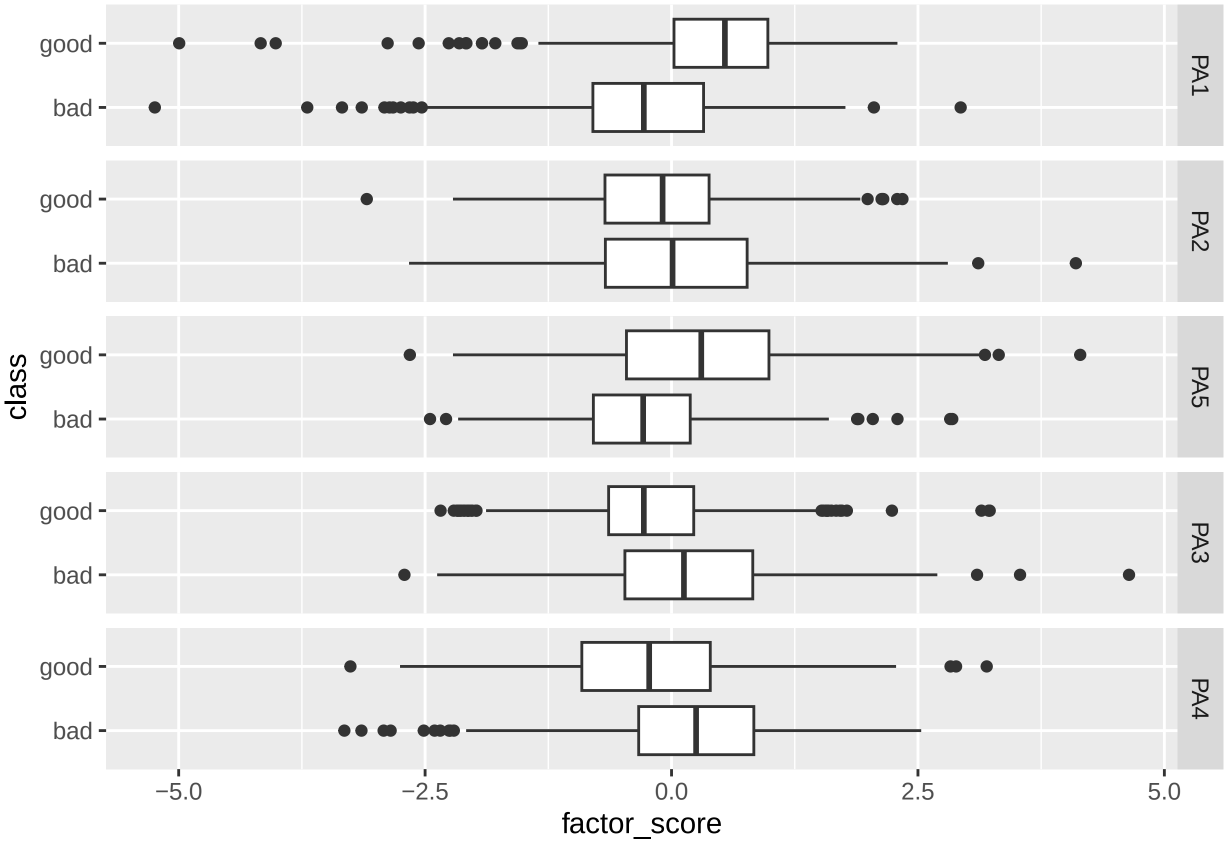
```
analyze_distributions <- function(data_factors_long, variable) {
  plot <- data_factors_long %>%
    ggplot(aes(x = factor_score, y = !!sym(variable))) +
    geom_boxplot() +
    facet_grid(factor ~ .)
  print(plot)

  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

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



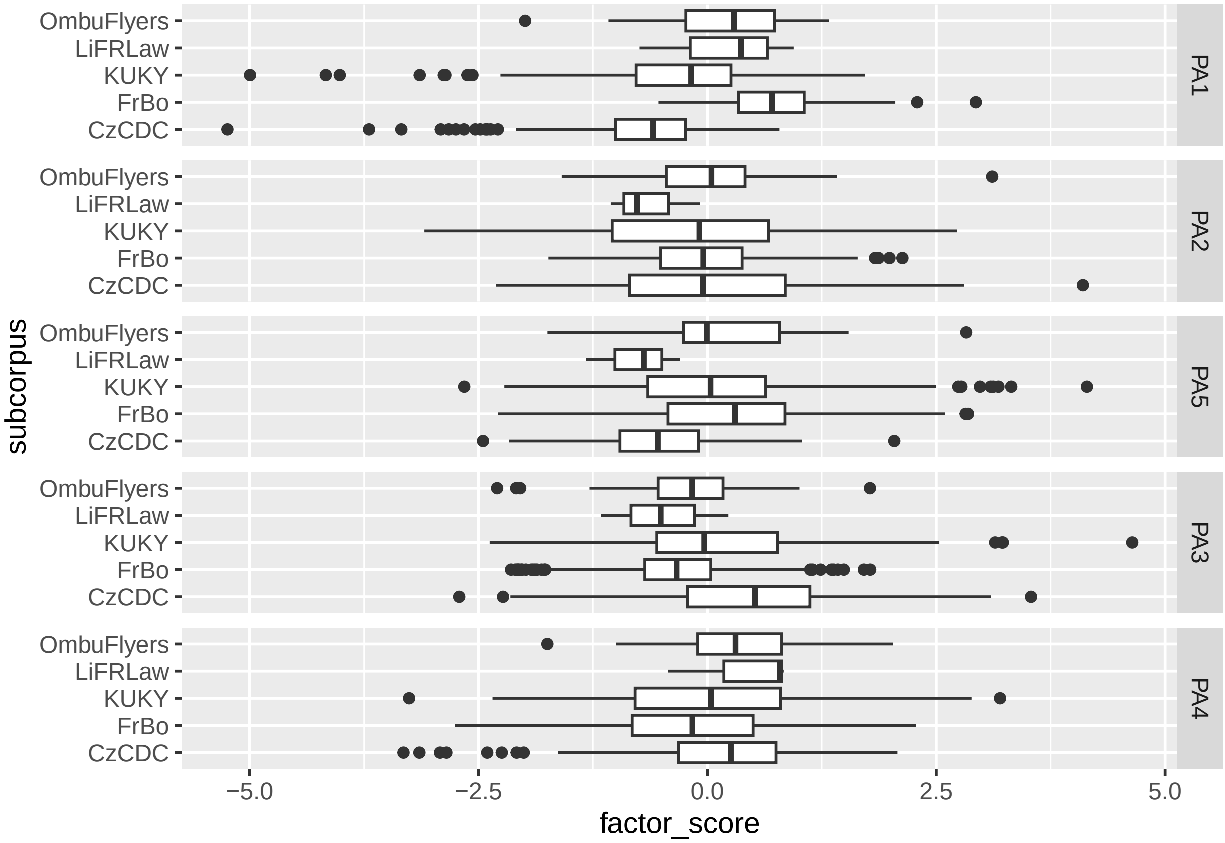
```
##
## Test for the significance of differences in class over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 123.8025, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## 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

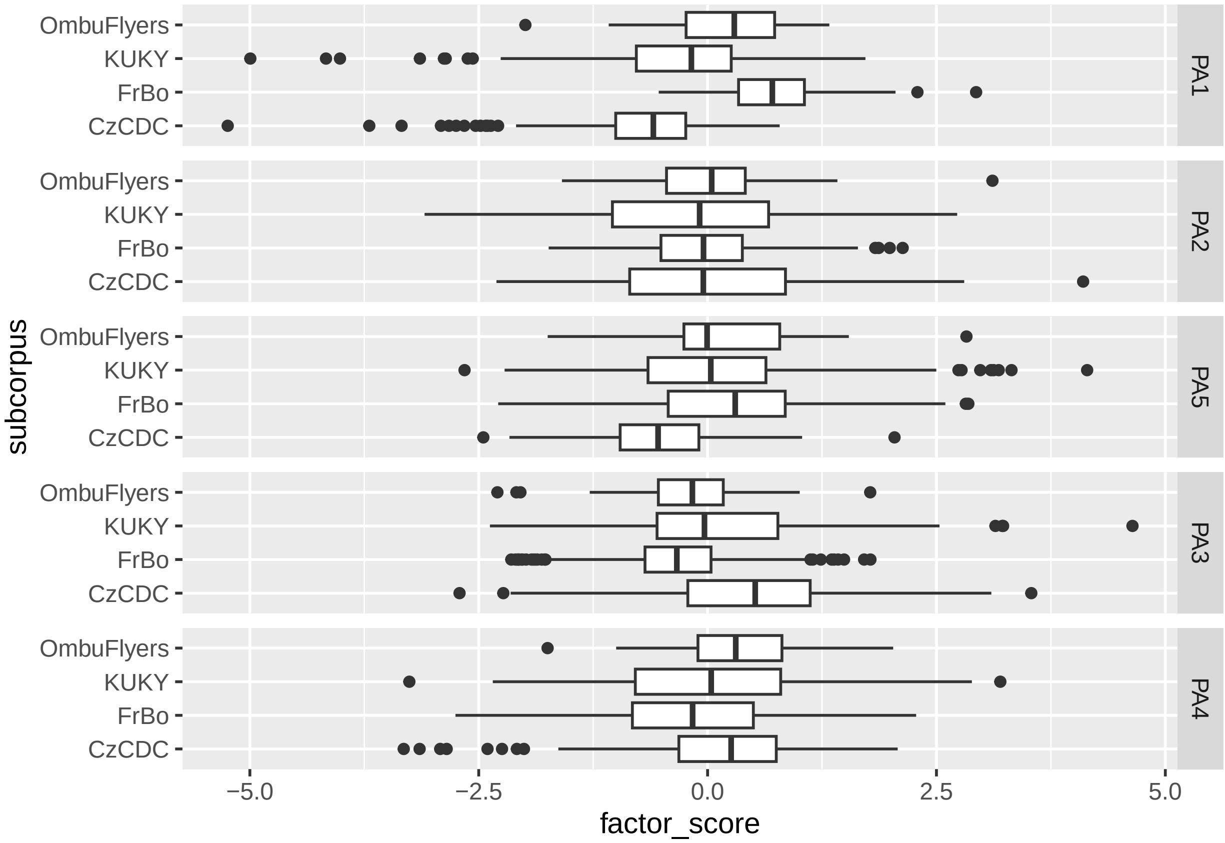
```
analyze_distributions(data_factors_long, "subcorpus")
```



```
##
## 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"
)
```



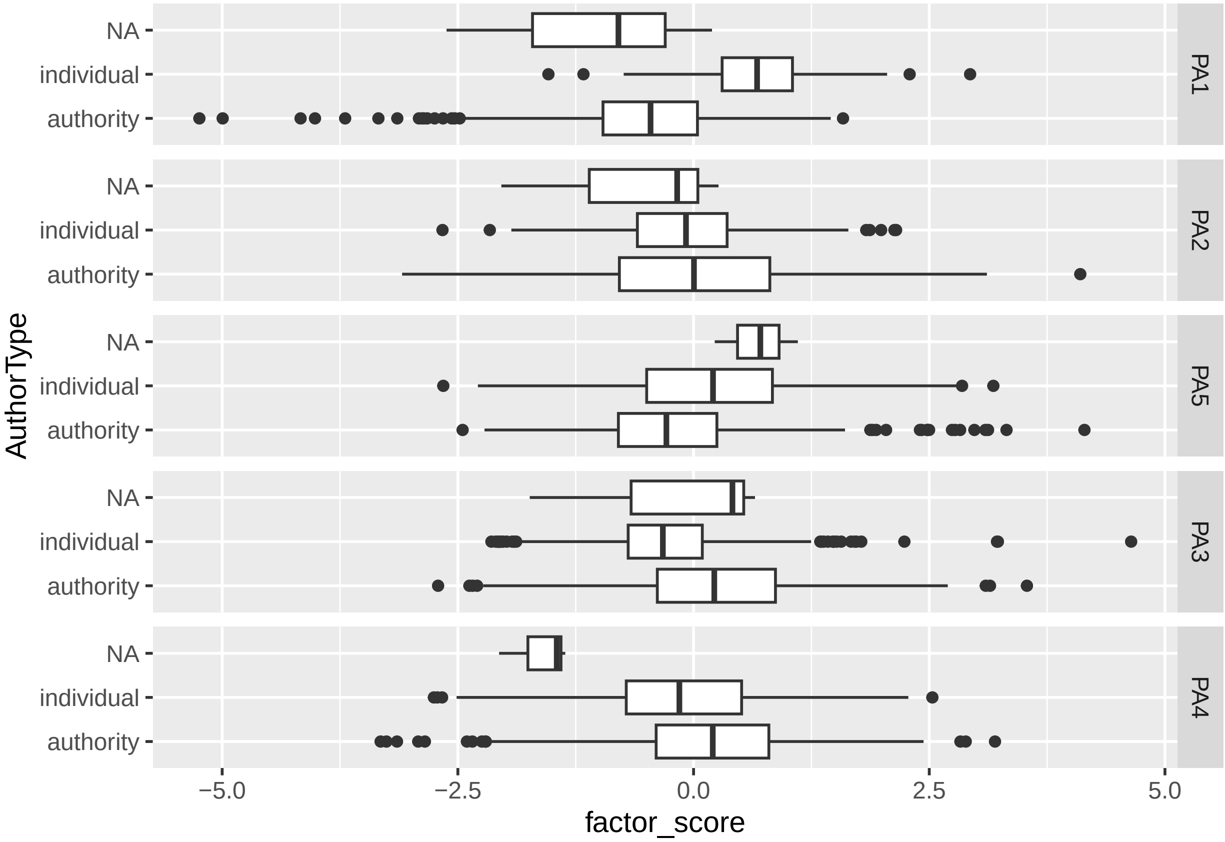
```
##
## Test for the significance of differences in subcorpus over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 363.4485, df = 3, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## 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")
```

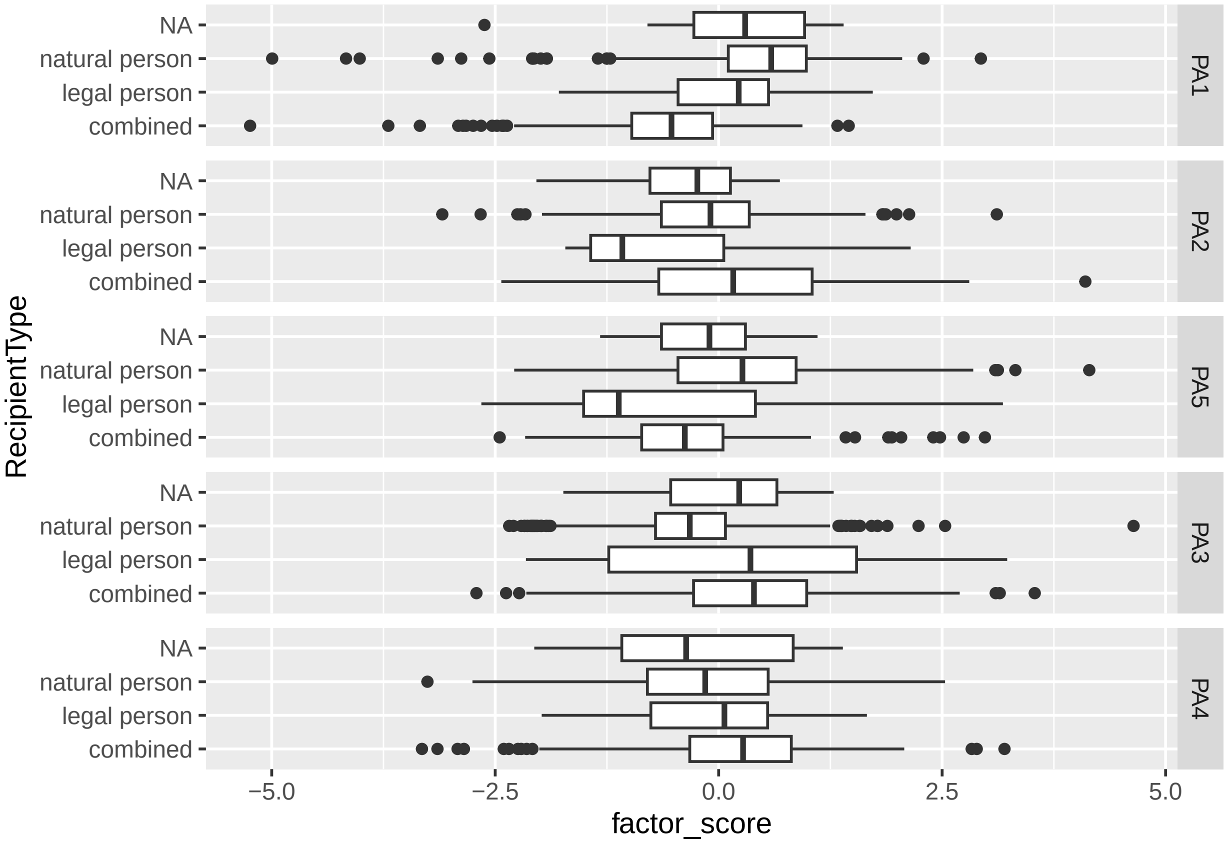


```
##
## Test for the significance of differences in AuthorType over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 337.0782, df = 1, p-value = 0
##
##
##                              Comparison of x by group
##                                    (Bonferroni)
## 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")
```



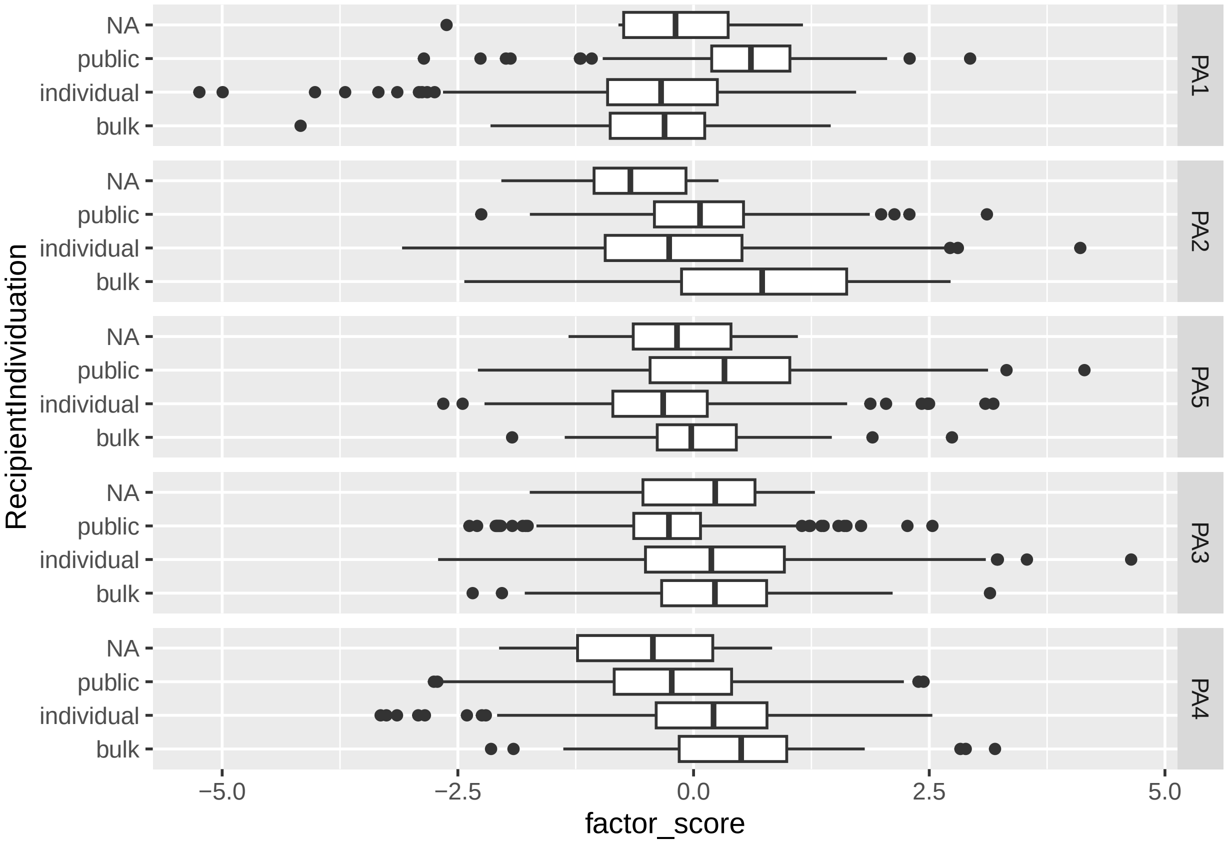
```
##
## Test for the significance of differences in RecipientType over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 272.2069, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## 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

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

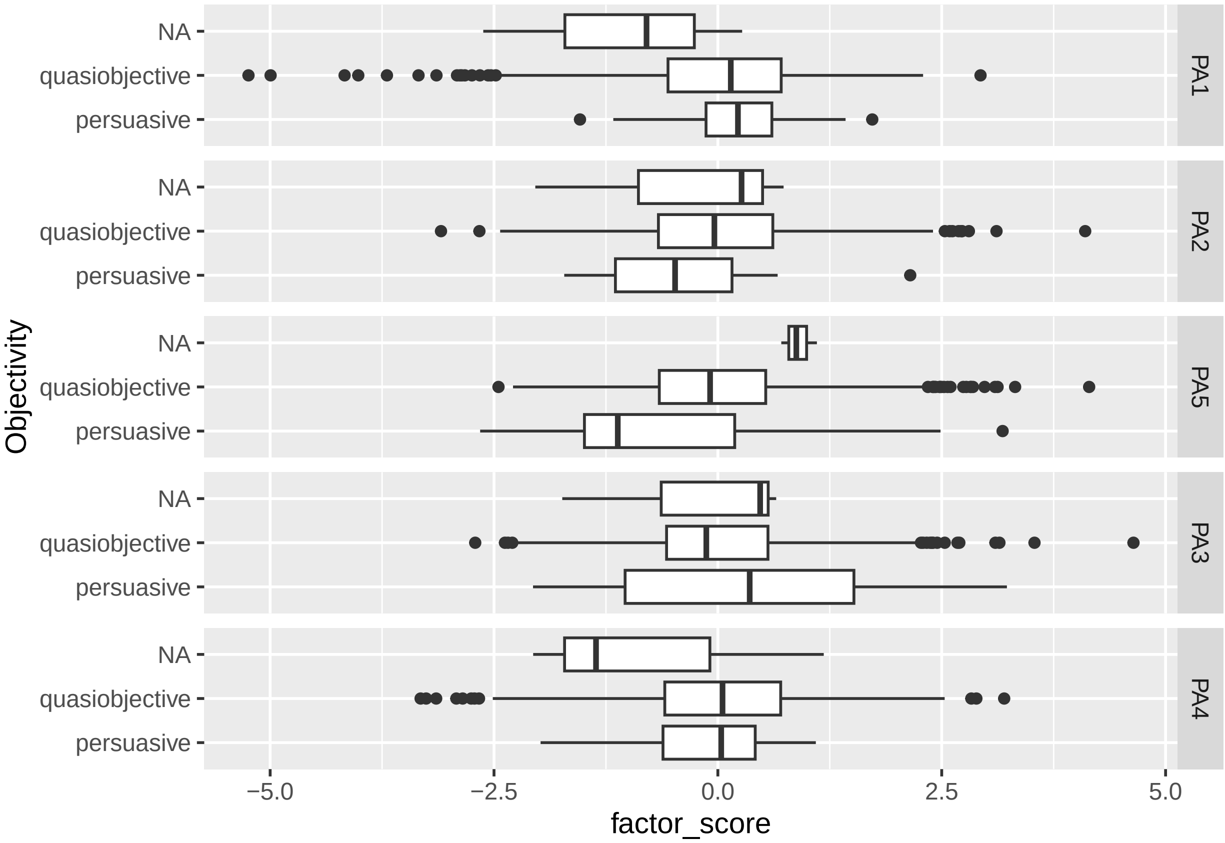


```
##
## Test for the significance of differences in RecipientIndividuation over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 210.8299, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (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

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



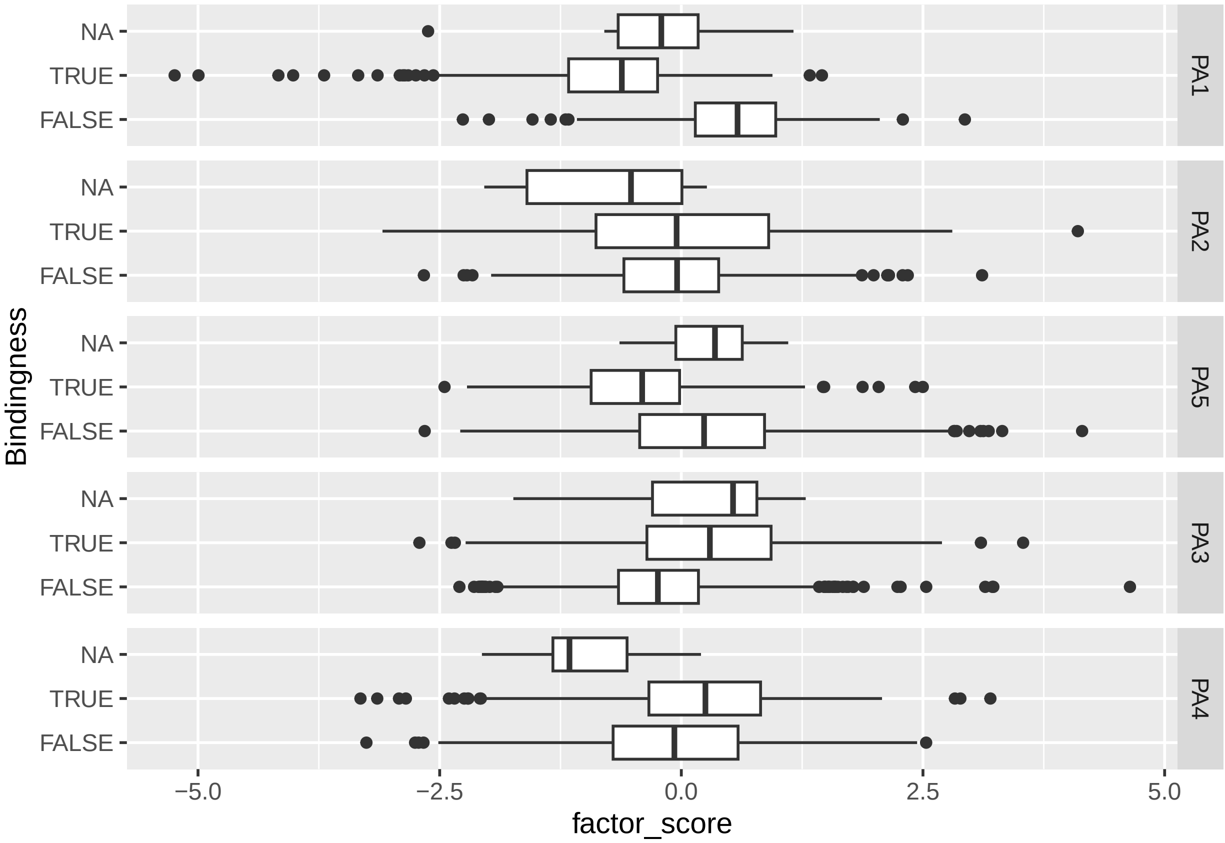
```
##
## Test for the significance of differences in Objectivity over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.5005, df = 1, p-value = 0.48
##
##
##                              Comparison of x by group
##                                    (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
##
```

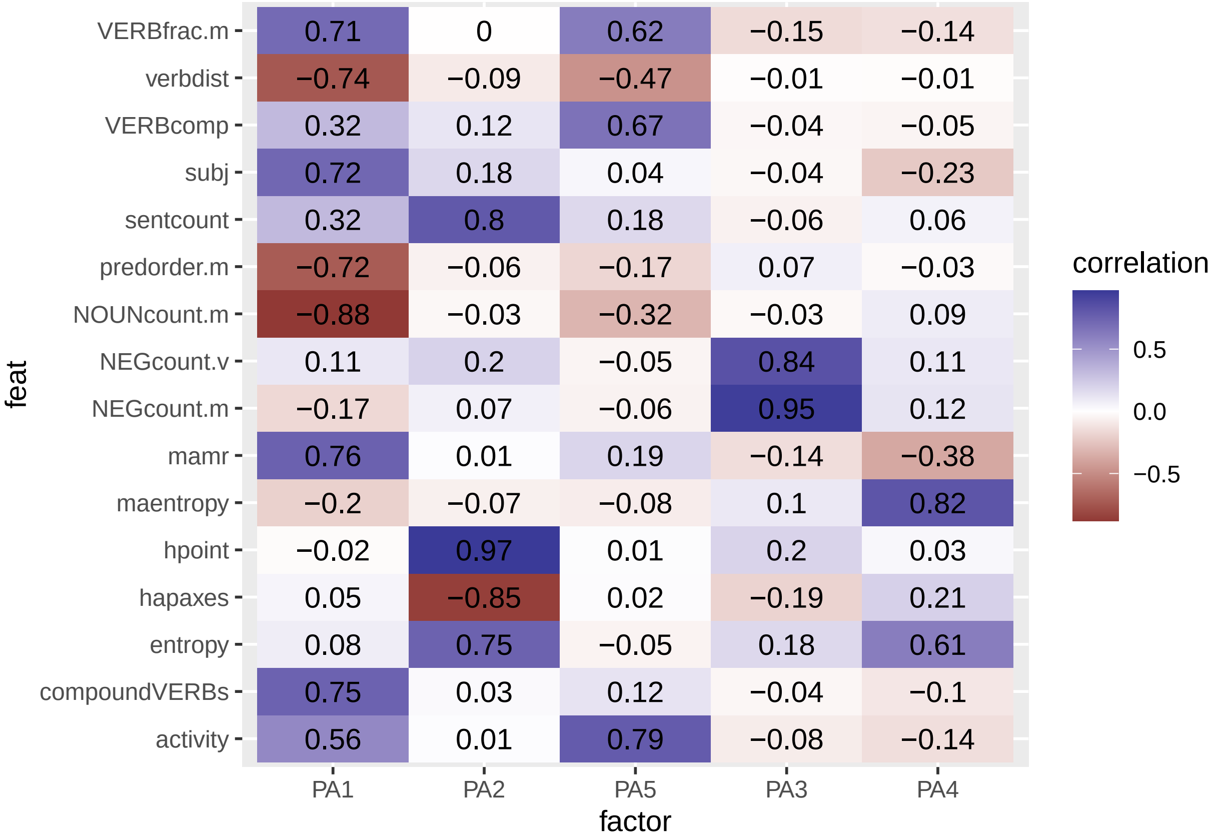
```
## Test for the significance of differences in Bindingness over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 16.5311, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## 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

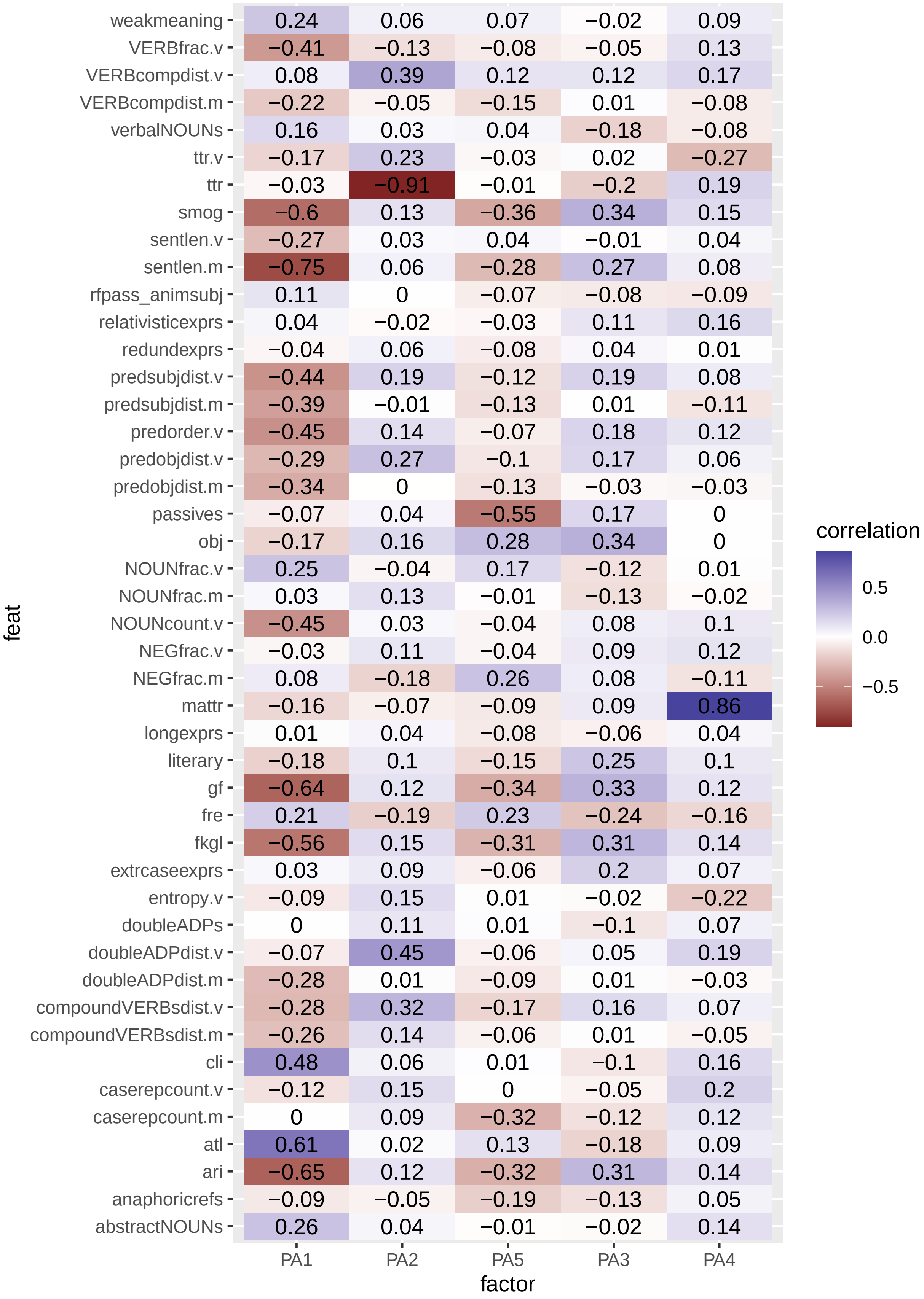
```
data_factors_longer <- data_factors_long %>%
  pivot_longer(
    abstractNOUNs:verbdist,
    names_to = "feat", values_to = "feat_value"
  )

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	PA1	PA2	PA5	PA3	PA4
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
doubleADPdist.v	−0.07	0.45	−0.06	0.05	0.19
doubleADPdist.m	−0.28	0.01	−0.09	0.01	−0.03
compoundVERBsdist.v	−0.28	0.32	−0.17	0.16	0.07
compoundVERBsdist.m	−0.26	0.14	−0.06	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