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
```

```r
library(paletteer) # color palettes

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

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

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

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

Load and tidy data

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

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

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

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


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

```
## Rows: 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.
```

```r
.firstnonmetacolumn <- 17

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

```
  mutate(
    GPs = RuleGPcoordovs +
      RuleGPdeverbaddr +
      RuleGPpatinstr +
      RuleGPdeverbsubj +
      RuleGPadjective +
      RuleGPpatbenperson +
      RuleGPwordorder
  ) %>%
  select(!c(
    RuleGPcoordovs,
    RuleGPdeverbaddr,
    RuleGPpatinstr,
    RuleGPdeverbsubj,
    RuleGPadjective,
    RuleGPpatbenperson,
    RuleGPwordorder
  ))

data_clean <- data_no_nas %>%
  # norm data expected to correlate with text length
  mutate(across(c(
    GPs,
    RuleDoubleAdpos,
    RuleAmbiguousRegards,
    RuleFunctionWordRepetition,
    RuleWeakMeaningWords,
    RuleAbstractNouns,
    RuleRelativisticExpressions,
    RuleConfirmationExpressions,
    RuleRedundantExpressions,
    RuleTooLongExpressions,
    RuleAnaphoricReferences,
    RuleLiteraryStyle,
    RulePassive,
    RuleVerbalNouns,
    RuleDoubleComparison,
    RuleWrongValencyCase,
    RuleWrongVerbonominalCase,
    RuleIncompleteConjunction,
    num_hapax,
    RuleReflexivePassWithAnimSubj,
    RuleTooManyNominalConstructions,
    RulePredSubjDistance,
    RuleMultiPartVerbs,
    RulePredAtClauseBeginning
  ), ~ .x / word_count)) %>%
  mutate(across(c(
    RuleTooFewVerbs,
    RuleTooManyNegations,
    RuleCaseRepetition,
    RuleLongSentences,
    RulePredObjDistance,
    RuleInfVerbDistance
  ), ~ .x / sent_count)) %>%
  # remove variables identified as text-length dependent
  select(!c(
    RuleTooFewVerbs,
    RuleTooManyNegations,
    RuleTooManyNominalConstructions,
    RuleCaseRepetition,
    RuleLongSentences,
    RulePredAtClauseBeginning,
    syllab_count,
    char_count
  )) %>%
  # remove variables identified as unreliable
  select(!c(
    RuleAmbiguousRegards,
    RuleFunctionWordRepetition,
    RuleDoubleComparison,
    RuleWrongValencyCase,
    RuleWrongVerbonominalCase
  )) %>%
  # remove further variables belonging to the 'acceptability' category
  select(!c(RuleIncompleteConjunction)) %>%
  # remove artificially limited variables
  select(!c(
    RuleCaseRepetition.max_repetition_frac,
    RuleCaseRepetition.max_repetition_frac.v
  )) %>%
  # remove variables with too many NAs
  select(!c(
    RuleDoubleAdpos.max_allowable_distance,
    RuleDoubleAdpos.max_allowable_distance.v
  )) %>%
  mutate(across(c(
    class,
    FileFormat,
    subcorpus,
    DocumentVersion,
    LegalActType,
    Objectivity,
    AuthorType,
    RecipientType,
    RecipientIndividuation,
    Anonymized
  ), ~ as.factor(.x)))

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

```
## # A tibble: 0 x 77
## # i 77 variables: KUK_ID <chr>, FileName <chr>, FileFormat <fct>,
## #   subcorpus <fct>, SourceID <chr>, DocumentVersion <fct>,
## #   ParentDocumentID <chr>, LegalActType <fct>, Objectivity <fct>,
## #   Bindingness <lgl>, AuthorType <fct>, RecipientType <fct>,
## #   RecipientIndividuation <fct>, Anonymized <fct>, Recipient Type <chr>,
## #   class <fct>, RuleAbstractNouns <dbl>, RuleAnaphoricReferences <dbl>,
## #   RuleCaseRepetition.max_repetition_count <dbl>, ...
```

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

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

Important features identification

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

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

  # formula_single <- reformulate(fname, "class")

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

  kw <- kruskal.test(data_clean[[i]], data_clean$class)
  p_value <- kw$p.value

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

```
## # A tibble: 61 x 2
##    feat_name                                  p_value
##    <chr>                                        <dbl>
##  1 RuleAbstractNouns                         6.39e- 3
##  2 RuleAnaphoricReferences                   9.79e- 3
##  3 RuleCaseRepetition.max_repetition_count   7.60e- 2
##  4 RuleCaseRepetition.max_repetition_count.v 9.43e- 4
##  5 RuleConfirmationExpressions               1.34e- 3
##  6 RuleDoubleAdpos                           3.02e- 1
##  7 RuleInfVerbDistance                       1.36e-16
##  8 RuleInfVerbDistance.max_distance          1.73e- 2
##  9 RuleInfVerbDistance.max_distance.v        7.89e- 2
## 10 RuleLiteraryStyle                         1.44e-26
## # i 51 more rows
```

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

Correlations

See Levshina (2015: 353–54).

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

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

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

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

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

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

what unites the low-communality variables we threw out:

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

High correlations

```
.hcorrcutoff <- 0.9

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

```
## # A tibble: 22 x 4
##    feat1                        feat2                          cor abs_cor
##    <chr>                        <chr>                        <dbl>   <dbl>
##  1 RuleLongSentences.max_length ari                          0.943   0.943
##  2 RuleLongSentences.max_length gf                           0.922   0.922
##  3 ari                          fkgl                         0.984   0.984
##  4 ari                          gf                           0.978   0.978
##  5 ari                          smog                         0.951   0.951
##  6 ari                          RuleLongSentences.max_length 0.943   0.943
##  7 atl                          cli                          0.960   0.960
##  8 cli                          atl                          0.960   0.960
##  9 fkgl                         ari                          0.984   0.984
## 10 fkgl                         gf                           0.967   0.967
## 11 fkgl                         smog                         0.948   0.948
## 12 gf                           smog                         0.987   0.987
## 13 gf                           ari                          0.978   0.978
## 14 gf                           fkgl                         0.967   0.967
## 15 gf                           RuleLongSentences.max_length 0.922   0.922
## 16 hpoint                       word_count                   0.958   0.958
## 17 maentropy                    mattr                        0.964   0.964
## 18 mattr                        maentropy                    0.964   0.964
## 19 smog                         gf                           0.987   0.987
## 20 smog                         ari                          0.951   0.951
## 21 smog                         fkgl                         0.948   0.948
## 22 word_count                   hpoint                       0.958   0.958
```

exclude:

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

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

```
## Compare row 8  and column  34 with corr  0.943
##   Means:  0.404 vs 0.213 so flagging column 8
## Compare row 34  and column  40 with corr  0.978
##   Means:  0.387 vs 0.205 so flagging column 34
## Compare row 40  and column  47 with corr  0.987
##   Means:  0.373 vs 0.198 so flagging column 40
## Compare row 47  and column  38 with corr  0.948
##   Means:  0.353 vs 0.191 so flagging column 47
## Compare row 35  and column  36 with corr  0.96
##   Means:  0.258 vs 0.186 so flagging column 35
## Compare row 49  and column  41 with corr  0.958
##   Means:  0.182 vs 0.183 so flagging column 41
## Compare row 42  and column  44 with corr  0.964
##   Means:  0.17 vs 0.184 so flagging column 44
## All correlations <= 0.9
```

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

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

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

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

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

Low correlations

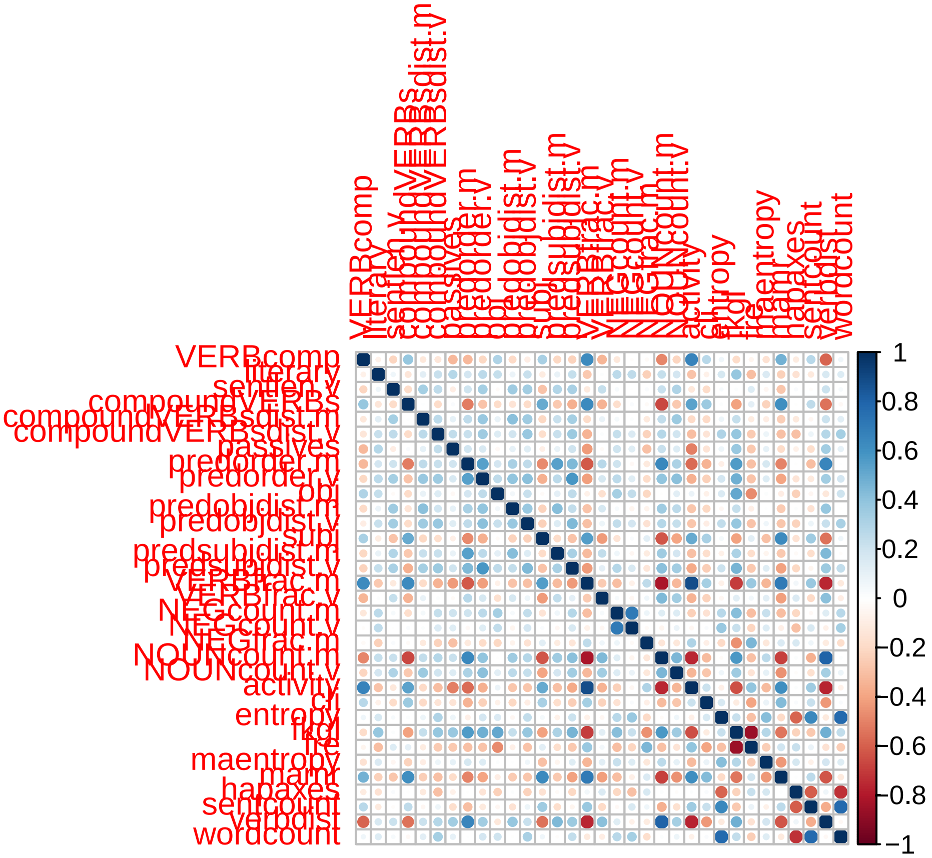
```r
# 0.35 instead of 0.3 otherwise the FA bootstrapping would freeze
.lcorrcutoff <- 0.35

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                               0.00639
## 2 RuleAnaphoricReferences                         0.00979
## 3 RuleCaseRepetition.max_repetition_count.v       0.000943
## 4 RuleConfirmationExpressions                     0.00134
## 5 RuleInfVerbDistance.max_distance                0.0173
## 6 RuleRedundantExpressions                        0.00129
## 7 RuleRelativisticExpressions                     0.0000178
## 8 RuleTooManyNominalConstructions.max_noun_frac.v 0.000000195
## 9 RuleVerbalNouns                                 0.000356
```

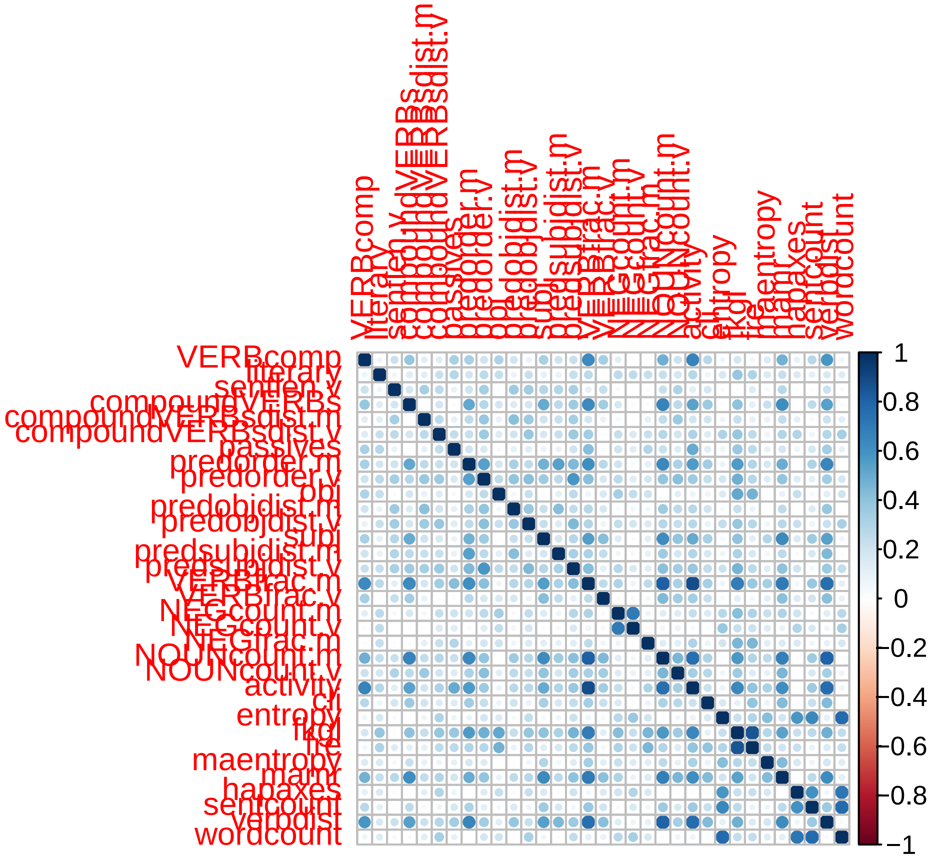
```
data_pure <- data_pureish_striphigh %>%
  select(!any_of(low_correlating_features))

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

colnames(data_pure) <- cnames
```

Visualisation

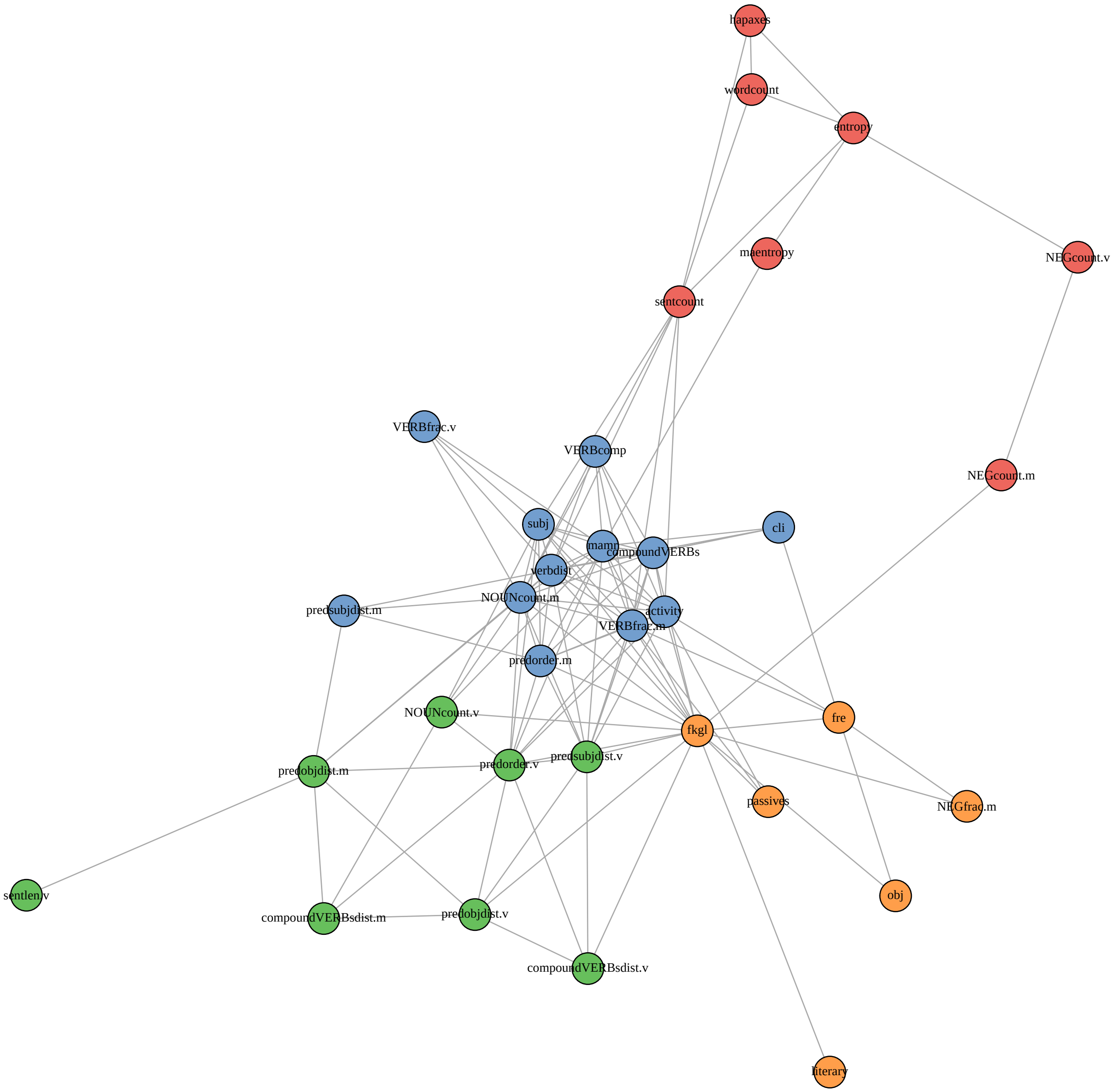
```
corrplot(cor(data_pure))
```



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



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

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

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

network_membership <- membership(network_communities)

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



Scaling

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

Check for normality

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

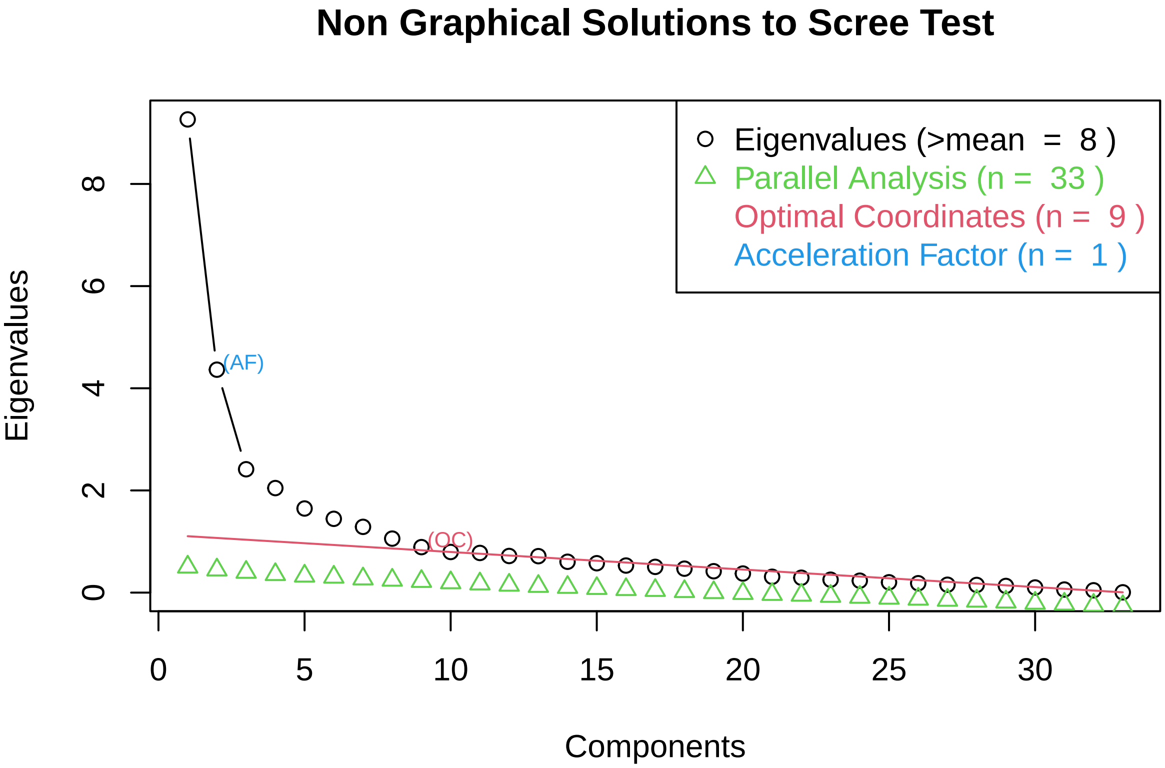
```
##          Beta-hat       kappa p-val
## Skewness 1054.115 132291.4622     0
## Kurtosis 2695.647    439.8094     0
```

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

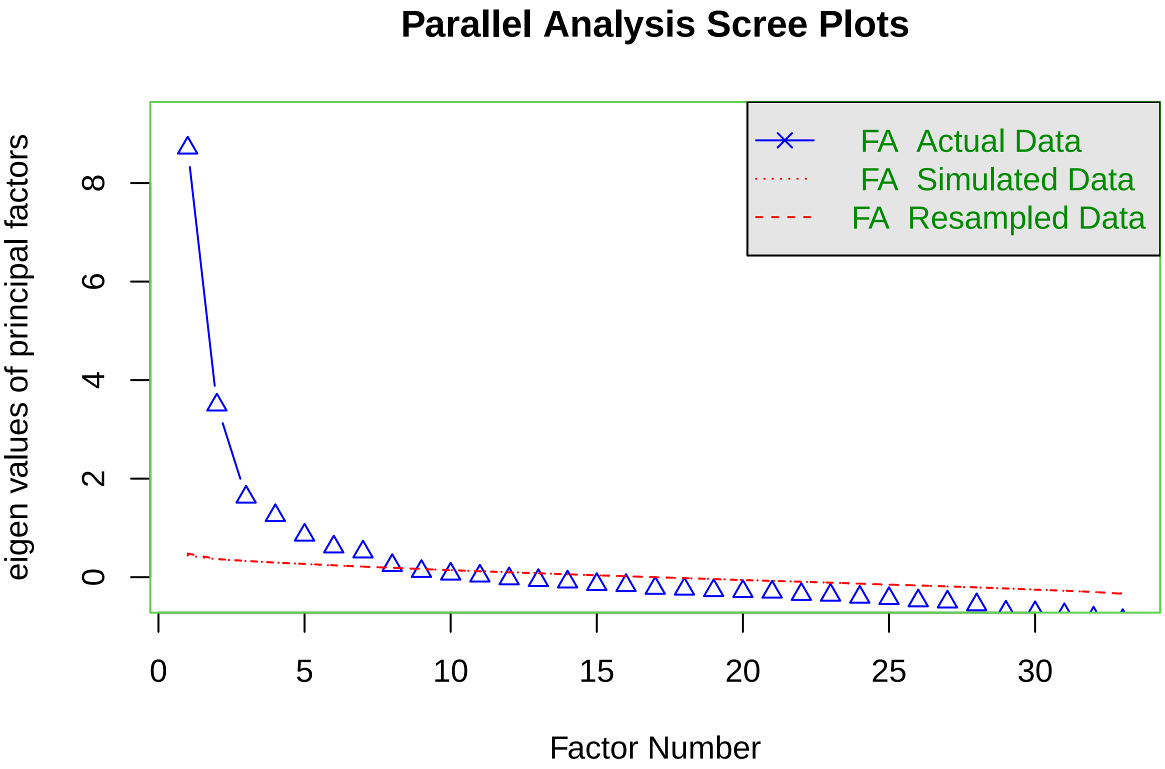
No. of factors

```
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 =  8  and the number of components =  NA
```

Model

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

```r
set.seed(42)

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.
```

```r
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   PA4   PA3   PA6   PA5   PA8   PA7   h2    u2
## VERBcomp             0.62  0.02 -0.01  0.54  0.27 -0.12 -0.02  0.04 0.60 0.404
## literary             0.03 -0.04  0.09  0.16 -0.29  0.14  0.07 -0.04 0.23 0.766
## sentlen.v            0.05 -0.01  0.78 -0.19  0.25  0.02  0.02  0.02 0.48 0.521
## compoundVERBs        0.96 -0.13  0.28 -0.26 -0.36  0.02 -0.07  0.15 0.70 0.296
## compoundVERBsdist.m  0.21 -0.02  0.73 -0.04 -0.15 -0.08 -0.10 -0.06 0.42 0.577
## compoundVERBsdist.v -0.08  0.24  0.31  0.01 -0.19  0.03 -0.04 -0.03 0.33 0.672
## passives            -0.02 -0.08 -0.03 -0.24 -0.86  0.09 -0.09 -0.12 0.58 0.419
## predorder.m         -0.67 -0.05  0.12  0.23  0.09 -0.01 -0.16  0.00 0.63 0.370
## predorder.v         -0.07 -0.01  0.56  0.18 -0.01  0.06  0.02 -0.02 0.53 0.474
## obj                  0.16 -0.05 -0.03  0.95  0.18  0.12 -0.10 -0.12 0.70 0.302
## predobjdist.m       -0.06 -0.09  0.64 -0.09  0.00 -0.04 -0.15  0.08 0.40 0.598
## predobjdist.v        0.04  0.14  0.55  0.06 -0.02  0.08  0.01  0.07 0.40 0.598
## subj                 0.51  0.14 -0.15 -0.06 -0.09  0.08 -0.31  0.09 0.57 0.431
## predsubjdist.m      -0.37 -0.02  0.32  0.07  0.13 -0.01 -0.30  0.12 0.39 0.607
## predsubjdist.v      -0.18  0.10  0.42  0.16 -0.01  0.09 -0.04 -0.05 0.46 0.536
## VERBfrac.m           0.87 -0.05  0.18  0.03  0.33 -0.02 -0.06  0.06 0.90 0.100
## VERBfrac.v          -0.47 -0.05  0.15 -0.19  0.21 -0.02  0.19  0.02 0.33 0.668
## NEGcount.m          -0.04 -0.09 -0.05  0.17  0.07  0.95  0.06  0.00 0.94 0.059
## NEGcount.v           0.21  0.06  0.03  0.06 -0.03  0.71  0.11  0.03 0.59 0.410
## NEGfrac.m           -0.08 -0.03 -0.06 -0.21  0.49  0.30 -0.12 -0.06 0.41 0.593
## NOUNcount.m         -0.90  0.03  0.03 -0.01  0.00 -0.12  0.09  0.03 0.82 0.184
## NOUNcount.v         -0.10 -0.07  0.43  0.06 -0.04 -0.03  0.16 -0.12 0.36 0.639
## activity             0.79 -0.01  0.08  0.27  0.46  0.00 -0.10 -0.10 0.92 0.080
## cli                  0.31 -0.02  0.02 -0.12  0.16  0.02  0.27  0.88 0.81 0.188
## entropy              0.04  0.75  0.07 -0.10  0.00  0.06  0.45  0.14 0.86 0.143
## fkgl                -0.41 -0.04 -0.05  0.57 -0.26  0.04  0.06  0.13 0.97 0.033
## fre                  0.13  0.04  0.06 -0.52  0.16 -0.05 -0.14 -0.58 0.97 0.034
## maentropy           -0.27  0.01 -0.15 -0.05  0.01  0.09  0.66  0.20 0.50 0.497
## mamr                 0.64 -0.05 -0.06 -0.04  0.01  0.00 -0.37  0.19 0.78 0.219
## hapaxes              0.07 -0.80  0.07 -0.13  0.06  0.00  0.24  0.14 0.70 0.304
## sentcount            0.12  0.98  0.01 -0.23  0.27 -0.08  0.00  0.07 0.93 0.068
## verbdist            -0.87  0.00  0.03 -0.21 -0.17 -0.05 -0.10 -0.06 0.81 0.192
## wordcount           -0.11  0.95  0.00 -0.02  0.01  0.00  0.07 -0.04 0.89 0.109
##                     com
## VERBcomp            2.5
## literary            2.6
## sentlen.v           1.3
## compoundVERBs       1.8
## compoundVERBsdist.m 1.3
## compoundVERBsdist.v 2.9
## passives            1.3
## predorder.m         1.5
## predorder.v         1.3
## obj                 1.2
## predobjdist.m       1.3
## predobjdist.v       1.3
## subj                2.3
## predsubjdist.m      3.6
## predsubjdist.v      2.0
## VERBfrac.m          1.4
## VERBfrac.v          2.5
## NEGcount.m          1.1
## NEGcount.v          1.3
## NEGfrac.m           2.4
## NOUNcount.m         1.1
## NOUNcount.v         1.7
## activity            2.0
## cli                 1.6
## entropy             1.8
## fkgl                2.5
## fre                 2.4
## maentropy           1.7
## mamr                1.9
## hapaxes             1.3
## sentcount           1.3
## verbdist            1.2
## wordcount           1.0
##
##                        PA1  PA2  PA4  PA3  PA6  PA5  PA8  PA7
## SS loadings           6.46 3.09 2.78 2.24 2.02 1.65 1.34 1.33
## Proportion Var        0.20 0.09 0.08 0.07 0.06 0.05 0.04 0.04
## Cumulative Var        0.20 0.29 0.37 0.44 0.50 0.55 0.59 0.63
## Proportion Explained  0.31 0.15 0.13 0.11 0.10 0.08 0.06 0.06
## Cumulative Proportion 0.31 0.46 0.59 0.70 0.79 0.87 0.94 1.00
##
##  With factor correlations of
##       PA1   PA2   PA4   PA3   PA6   PA5   PA8   PA7
## PA1  1.00  0.11 -0.59 -0.28  0.38 -0.21 -0.16  0.07
## PA2  0.11  1.00  0.15  0.31 -0.27  0.31  0.09  0.16
## PA4 -0.59  0.15  1.00  0.38 -0.32  0.22  0.15 -0.12
## PA3 -0.28  0.31  0.38  1.00 -0.48  0.26  0.17  0.22
## PA6  0.38 -0.27 -0.32 -0.48  1.00 -0.29 -0.14 -0.29
## PA5 -0.21  0.31  0.22  0.26 -0.29  1.00  0.15 -0.05
## PA8 -0.16  0.09  0.15  0.17 -0.14  0.15  1.00 -0.18
## PA7  0.07  0.16 -0.12  0.22 -0.29 -0.05 -0.18  1.00
##
## Mean item complexity =  1.8
## Test of the hypothesis that 8 factors are sufficient.
##
## df null model =  528  with the objective function =  27.53 with Chi Square =  20379.34
## df of  the model are 292  and the objective function was  3.91
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.03
##
## The harmonic n.obs is  753 with the empirical chi square  501.9  with prob <  2.6e-13
## The total n.obs was  753  with Likelihood Chi Square =  2874.96  with prob <  0
##
## Tucker Lewis Index of factoring reliability =  0.763
## RMSEA index =  0.108  and the 90 % confidence intervals are  0.105 0.112
## BIC =  940.73
## Fit based upon off diagonal values = 0.99
##  Coefficients and bootstrapped confidence intervals
##                       low   PA1 upper   low   PA2 upper   low   PA4 upper   low
## VERBcomp             0.49  0.62  0.72 -0.03  0.02  0.07 -0.07 -0.01  0.06  0.43
## literary            -0.08  0.03  0.11 -0.11 -0.04  0.03 -0.01  0.09  0.18  0.05
## sentlen.v           -0.05  0.05  0.11 -0.07 -0.01  0.06  0.57  0.78  0.91 -0.26
## compoundVERBs        0.73  0.96  1.10 -0.18 -0.13 -0.07  0.15  0.28  0.36 -0.33
## compoundVERBsdist.m  0.05  0.21  0.33 -0.10 -0.02  0.06  0.54  0.73  0.87 -0.11
## compoundVERBsdist.v -0.19 -0.08  0.01  0.17  0.24  0.32  0.16  0.31  0.44 -0.06
## passives            -0.12 -0.02  0.03 -0.12 -0.08 -0.03 -0.09 -0.03  0.03 -0.32
## predorder.m         -0.75 -0.67 -0.52 -0.10 -0.05  0.00  0.01  0.12  0.24  0.12
## predorder.v         -0.23 -0.07  0.07 -0.08 -0.01  0.07  0.33  0.56  0.76  0.09
## obj                  0.07  0.16  0.22 -0.10 -0.05 -0.01 -0.08 -0.03  0.04  0.81
## predobjdist.m       -0.24 -0.06  0.13 -0.17 -0.09 -0.03  0.45  0.64  0.81 -0.16
## predobjdist.v       -0.10  0.04  0.18  0.04  0.14  0.24  0.40  0.55  0.67 -0.04
## subj                 0.42  0.51  0.61  0.08  0.14  0.19 -0.22 -0.15 -0.08 -0.13
## predsubjdist.m      -0.48 -0.37 -0.24 -0.06 -0.02  0.03  0.20  0.32  0.43 -0.06
## predsubjdist.v      -0.31 -0.18 -0.06  0.04  0.10  0.18  0.25  0.42  0.54  0.08
## VERBfrac.m           0.71  0.87  0.99 -0.08 -0.05 -0.01  0.10  0.18  0.23 -0.01
## VERBfrac.v          -0.59 -0.47 -0.32 -0.13 -0.05  0.02  0.01  0.15  0.28 -0.29
## NEGcount.m          -0.08 -0.04  0.04 -0.13 -0.09 -0.03 -0.10 -0.05  0.01  0.13
## NEGcount.v           0.15  0.21  0.27  0.00  0.06  0.12 -0.03  0.03  0.08  0.01
## NEGfrac.m           -0.15 -0.08  0.06 -0.11 -0.03  0.03 -0.14 -0.06  0.04 -0.28
## NOUNcount.m         -1.01 -0.90 -0.71 -0.02  0.03  0.07 -0.03  0.03  0.10 -0.08
## NOUNcount.v         -0.25 -0.10  0.01 -0.16 -0.07  0.02  0.27  0.43  0.56 -0.02
## activity             0.65  0.79  0.89 -0.04 -0.01  0.02  0.03  0.08  0.13  0.21
## cli                  0.26  0.31  0.41 -0.06 -0.02  0.02 -0.06  0.02  0.07 -0.17
## entropy             -0.05  0.04  0.08  0.72  0.75  0.79  0.01  0.07  0.11 -0.16
## fkgl                -0.47 -0.41 -0.32 -0.07 -0.04 -0.02 -0.09 -0.05  0.01  0.49
## fre                  0.04  0.13  0.18  0.01  0.04  0.08  0.00  0.06  0.10 -0.63
## maentropy           -0.39 -0.27 -0.18 -0.04  0.01  0.08 -0.25 -0.15 -0.04 -0.14
## mamr                 0.55  0.64  0.76 -0.10 -0.05 -0.01 -0.13 -0.06  0.03 -0.09
## hapaxes             -0.02  0.07  0.12 -0.84 -0.80 -0.75 -0.02  0.07  0.11 -0.19
## sentcount            0.10  0.12  0.19  0.92  0.98  1.02 -0.02  0.01  0.05 -0.28
## verbdist            -0.95 -0.87 -0.73 -0.03  0.00  0.02 -0.02  0.03  0.08 -0.25
## wordcount           -0.14 -0.11 -0.07  0.92  0.95  0.99 -0.03  0.00  0.04 -0.05
##                       PA3 upper   low   PA6 upper   low   PA5 upper   low   PA8
## VERBcomp             0.54  0.66  0.18  0.27  0.39 -0.19 -0.12 -0.04 -0.12 -0.02
## literary             0.16  0.26 -0.39 -0.29 -0.17  0.05  0.14  0.29 -0.03  0.07
## sentlen.v           -0.19 -0.11  0.16  0.25  0.34 -0.06  0.02  0.09 -0.08  0.02
## compoundVERBs       -0.26 -0.18 -0.47 -0.36 -0.22 -0.03  0.02  0.11 -0.22 -0.07
## compoundVERBsdist.m -0.04  0.03 -0.25 -0.15 -0.06 -0.18 -0.08 -0.01 -0.23 -0.10
## compoundVERBsdist.v  0.01  0.10 -0.29 -0.19 -0.08 -0.05  0.03  0.12 -0.16 -0.04
## passives            -0.24 -0.16 -0.97 -0.86 -0.72  0.01  0.09  0.22 -0.24 -0.09
## predorder.m          0.23  0.32 -0.09  0.09  0.20 -0.10 -0.01  0.09 -0.35 -0.16
## predorder.v          0.18  0.28 -0.12 -0.01  0.08 -0.01  0.06  0.16 -0.09  0.02
## obj                  0.95  1.09  0.11  0.18  0.27  0.05  0.12  0.25 -0.21 -0.10
## predobjdist.m       -0.09 -0.01 -0.19  0.00  0.17 -0.20 -0.04  0.07 -0.37 -0.15
## predobjdist.v        0.06  0.18 -0.15 -0.02  0.09 -0.03  0.08  0.17 -0.12  0.01
## subj                -0.06  0.01 -0.19 -0.09 -0.01  0.00  0.08  0.15 -0.48 -0.31
## predsubjdist.m       0.07  0.21 -0.01  0.13  0.28 -0.10 -0.01  0.10 -0.56 -0.30
## predsubjdist.v       0.16  0.25 -0.12 -0.01  0.10 -0.01  0.09  0.21 -0.20 -0.04
## VERBfrac.m           0.03  0.08  0.24  0.33  0.44 -0.07 -0.02  0.04 -0.17 -0.06
## VERBfrac.v          -0.19 -0.08  0.09  0.21  0.34 -0.13 -0.02  0.11  0.04  0.19
## NEGcount.m           0.17  0.25 -0.05  0.07  0.13  0.81  0.95  1.20 -0.01  0.06
## NEGcount.v           0.06  0.12 -0.12 -0.03  0.05  0.56  0.71  1.05  0.03  0.11
## NEGfrac.m           -0.21 -0.11  0.35  0.49  0.56  0.18  0.30  0.40 -0.19 -0.12
## NOUNcount.m         -0.01  0.06 -0.10  0.00  0.06 -0.26 -0.12 -0.05  0.02  0.09
## NOUNcount.v          0.06  0.14 -0.14 -0.04  0.09 -0.14 -0.03  0.09  0.00  0.16
## activity             0.27  0.32  0.37  0.46  0.59 -0.03  0.00  0.06 -0.23 -0.10
## cli                 -0.12 -0.04 -0.01  0.16  0.23 -0.13  0.02  0.10  0.17  0.27
## entropy             -0.10 -0.04 -0.06  0.00  0.12  0.00  0.06  0.17  0.32  0.45
## fkgl                 0.57  0.67 -0.33 -0.26 -0.20  0.01  0.04  0.10  0.01  0.06
## fre                 -0.52 -0.44  0.10  0.16  0.29 -0.10 -0.05  0.02 -0.34 -0.14
## maentropy           -0.05  0.03 -0.10  0.01  0.18 -0.01  0.09  0.25  0.50  0.66
## mamr                -0.04  0.01 -0.10  0.01  0.05 -0.10  0.00  0.06 -0.57 -0.37
## hapaxes             -0.13 -0.07  0.00  0.06  0.14 -0.06  0.00  0.07  0.15  0.24
## sentcount           -0.23 -0.18  0.19  0.27  0.32 -0.16 -0.08 -0.04 -0.04  0.00
## verbdist            -0.21 -0.16 -0.33 -0.17 -0.06 -0.13 -0.05  0.00 -0.17 -0.10
## wordcount           -0.02  0.02 -0.03  0.01  0.05 -0.03  0.00  0.04  0.04  0.07
##                     upper   low   PA7 upper
## VERBcomp             0.03 -0.07  0.04  0.20
## literary             0.15 -0.13 -0.04  0.09
## sentlen.v            0.12 -0.10  0.02  0.12
## compoundVERBs        0.01  0.00  0.15  0.42
## compoundVERBsdist.m -0.01 -0.18 -0.06  0.04
## compoundVERBsdist.v  0.06 -0.14 -0.03  0.07
## passives            -0.01 -0.22 -0.12  0.00
## predorder.m          0.03 -0.30  0.00  0.19
## predorder.v          0.11 -0.18 -0.02  0.10
## obj                 -0.03 -0.21 -0.12 -0.04
## predobjdist.m        0.04 -0.08  0.08  0.29
## predobjdist.v        0.14 -0.06  0.07  0.22
## subj                -0.21  0.01  0.09  0.18
## predsubjdist.m      -0.14 -0.08  0.12  0.38
## predsubjdist.v       0.10 -0.24 -0.05  0.07
## VERBfrac.m           0.00 -0.04  0.06  0.19
## VERBfrac.v           0.41 -0.12  0.02  0.16
## NEGcount.m           0.20 -0.16  0.00  0.08
## NEGcount.v           0.25 -0.06  0.03  0.14
## NEGfrac.m           -0.01 -0.36 -0.06  0.07
## NOUNcount.m          0.24 -0.10  0.03  0.11
## NOUNcount.v          0.34 -0.29 -0.12  0.03
## activity            -0.05 -0.17 -0.10 -0.06
## cli                  0.53  0.74  0.88  1.32
## entropy              0.70  0.02  0.14  0.41
## fkgl                 0.15  0.08  0.13  0.28
## fre                 -0.05 -0.96 -0.58 -0.45
## maentropy            1.03  0.03  0.20  0.57
## mamr                -0.23  0.09  0.19  0.31
## hapaxes              0.38  0.04  0.14  0.32
## sentcount            0.07  0.00  0.07  0.12
## verbdist            -0.02 -0.21 -0.06  0.01
## wordcount            0.13 -0.10 -0.04  0.00
##
##  Interfactor correlations and bootstrapped confidence intervals
##          lower estimate upper
## PA1-PA2 -0.546    0.112  0.38
## PA1-PA4 -0.852   -0.589  0.33
## PA1-PA3 -0.967   -0.278  0.65
## PA1-PA6 -0.685    0.376  0.64
## PA1-PA5 -0.661   -0.214  0.34
## PA1-PA8 -0.499   -0.161  0.30
## PA1-PA7 -0.312    0.070  0.34
## PA2-PA4 -0.042    0.154  0.47
## PA2-PA3 -0.393    0.307  0.64
## PA2-PA6 -0.478   -0.268  0.66
## PA2-PA5 -0.238    0.312  0.61
## PA2-PA8 -0.293    0.092  0.48
## PA2-PA7 -0.360    0.163  0.32
## PA4-PA3 -0.472    0.377  0.69
## PA4-PA6 -0.620   -0.321  0.67
## PA4-PA5 -0.302    0.222  0.59
## PA4-PA8 -0.331    0.146  0.44
## PA4-PA7 -0.330   -0.121  0.32
## PA3-PA6 -0.822   -0.483  0.50
## PA3-PA5 -0.563    0.260  0.59
## PA3-PA8 -0.477    0.175  0.53
## PA3-PA7 -0.393    0.225  0.27
## PA6-PA5 -0.454   -0.293  0.45
## PA6-PA8 -0.448   -0.145  0.43
## PA6-PA7 -0.408   -0.289  0.28
## PA5-PA8 -0.364    0.146  0.33
## PA5-PA7 -0.398   -0.053  0.28
## PA8-PA7 -0.464   -0.181  0.22
```

Healthiness diagnostics

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

```
## # A tibble: 33 x 2
##    feat                maxload
##    <chr>                 <dbl>
##  1 literary              0.287
##  2 compoundVERBsdist.v   0.310
##  3 predsubjdist.m        0.369
##  4 predsubjdist.v        0.418
##  5 NOUNcount.v           0.427
##  6 VERBfrac.v            0.466
##  7 NEGfrac.m             0.489
##  8 subj                  0.514
##  9 predobjdist.v         0.549
## 10 predorder.v           0.565
## # i 23 more rows
```

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

```
##            literary compoundVERBsdist.v          VERBfrac.v         NOUNcount.v
##           0.2341369           0.3283465           0.3320223           0.3612375
##      predsubjdist.m       predobjdist.m       predobjdist.v           NEGfrac.m
##           0.3934739           0.4021256           0.4021592           0.4070422
## compoundVERBsdist.m      predsubjdist.v           sentlen.v           maentropy
##           0.4228078           0.4635282           0.4793816           0.5029226
##         predorder.v                subj            passives          NEGcount.v
##           0.5262035           0.5685364           0.5806282           0.5898322
##            VERBcomp         predorder.m             hapaxes                 obj
##           0.5957954           0.6300904           0.6964991           0.6984773
##       compoundVERBs                mamr            verbdist                 cli
##           0.7035553           0.7808688           0.8077875           0.8120289
##         NOUNcount.m             entropy           wordcount          VERBfrac.m
##           0.8160658           0.8572308           0.8905290           0.8997511
##            activity           sentcount          NEGcount.m                 fre
##           0.9201307           0.9315112           0.9413964           0.9664674
##                fkgl
##           0.9672468
```

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

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

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

```
##           wordcount         NOUNcount.m          NEGcount.m                 obj
##            1.043638            1.067998            1.110702            1.225188
##            verbdist       predobjdist.v            passives       predobjdist.m
##            1.239877            1.252131            1.258336            1.263412
##         predorder.v          NEGcount.v           sentcount           sentlen.v
##            1.271268            1.273698            1.326526            1.336486
## compoundVERBsdist.m             hapaxes          VERBfrac.m         predorder.m
##            1.344661            1.347024            1.401879            1.494645
##                 cli           maentropy         NOUNcount.v       compoundVERBs
##            1.577941            1.730967            1.742690            1.769526
##             entropy                mamr            activity      predsubjdist.v
##            1.804342            1.856131            1.987791            2.033363
##                subj           NEGfrac.m                 fre            VERBcomp
##            2.292316            2.354448            2.434264            2.459942
##          VERBfrac.v                fkgl            literary compoundVERBsdist.v
##            2.463680            2.479479            2.637398            2.887351
##      predsubjdist.m
##            3.570794
```

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

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

Feature engineering

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

det(cor(data_engineered_1))
```

```
## [1] 2.394366e-07
```

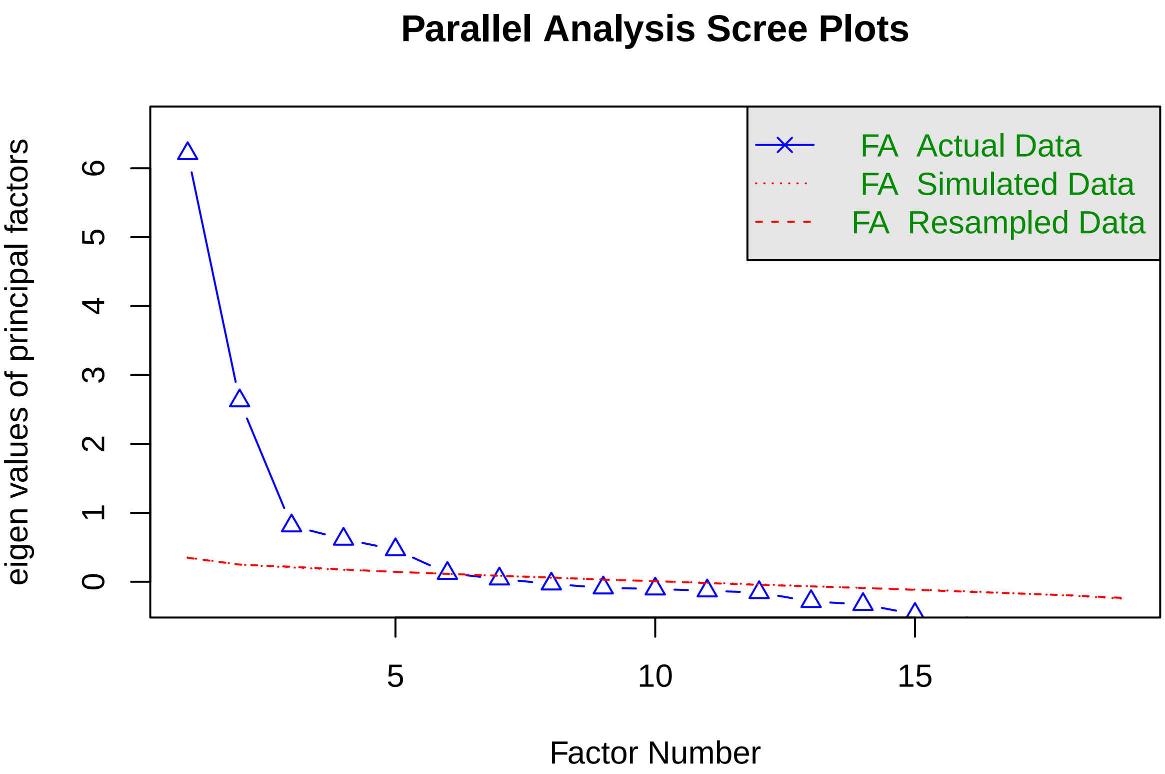
```
KMO(data_engineered_1)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = data_engineered_1)
## Overall MSA =  0.82
## MSA for each item = 
##      VERBcomp compoundVERBs      passives   predorder.m   predorder.v 
##          0.86          0.90          0.77          0.87          0.82 
##           obj          subj    VERBfrac.m    NEGcount.m    NEGcount.v 
##          0.50          0.93          0.88          0.72          0.67 
##   NOUNcount.m      activity       entropy     maentropy          mamr 
##          0.91          0.89          0.70          0.60          0.91 
##       hapaxes     sentcount      verbdist     wordcount 
##          0.78          0.69          0.92          0.69
```

second FA

No. of vectors

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



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

Model

```
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         0.26  0.05  0.59  0.05 -0.03 0.56 0.44 1.4
## compoundVERBs    0.80 -0.01 -0.14  0.10 -0.08 0.56 0.44 1.1
## passives         0.03  0.01 -0.59  0.23 -0.10 0.35 0.65 1.4
## predorder.m     -0.79 -0.03  0.01 -0.01 -0.15 0.60 0.40 1.1
## predorder.v     -0.53  0.10  0.05  0.15 -0.06 0.34 0.66 1.3
## obj             -0.29  0.00  0.45  0.41 -0.11 0.47 0.53 2.8
## subj           0.67  0.13 -0.08  0.06 -0.20 0.52 0.48 1.3
## VERBfrac.m     0.70 -0.04  0.40 -0.07 -0.02 0.89 0.11 1.6
## NEGcount.m     0.03 -0.10 -0.16  0.90  0.12 0.75 0.25 1.1
## NEGcount.v     0.27  0.04 -0.18  0.81  0.12 0.62 0.38 1.4
## NOUNcount.m   -0.87  0.04 -0.14 -0.18  0.01 0.82 0.18 1.1
## activity       0.54 -0.05  0.59 -0.02 -0.03 0.89 0.11 2.0
## entropy        0.03  0.77  0.03  0.13  0.44 0.87 0.13 1.7
## maentropy     -0.15  0.00  0.07  0.14  0.73 0.59 0.41 1.2
## mamr           0.71 -0.03  0.01 -0.04 -0.31 0.71 0.29 1.4
## hapaxes        0.11 -0.80  0.07 -0.04  0.31 0.73 0.27 1.4
## sentcount      0.24  0.90  0.09 -0.23  0.03 0.87 0.13 1.3
## verbdist      -0.70 -0.01 -0.37 -0.15 -0.08 0.77 0.23 1.7
## wordcount     -0.12  0.94 -0.03  0.02  0.04 0.89 0.11 1.0
##
##                        PA1  PA2  PA4  PA3  PA5
## SS loadings           5.13 2.94 1.92 1.74 1.08
## Proportion Var        0.27 0.15 0.10 0.09 0.06
## Cumulative Var        0.27 0.42 0.53 0.62 0.67
## Proportion Explained  0.40 0.23 0.15 0.14 0.08
## Cumulative Proportion 0.40 0.63 0.78 0.92 1.00
##
##  With factor correlations of
##       PA1  PA2   PA4   PA3   PA5
## PA1  1.00 0.07  0.38 -0.26 -0.20
## PA2  0.07 1.00  0.11  0.38  0.01
## PA4  0.38 0.11  1.00  0.08 -0.28
## PA3 -0.26 0.38  0.08  1.00 -0.04
## PA5 -0.20 0.01 -0.28 -0.04  1.00
##
## Mean item complexity =  1.4
## Test of the hypothesis that 5 factors are sufficient.
##
## df null model =  171  with the objective function =  15.24 with Chi Square =  11354.97
## df of  the model are 86  and the objective function was  1.78
##
## 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  279.26  with prob <  2.8e-22
## The total n.obs was  753  with Likelihood Chi Square =  1318.26  with prob <  1.1e-219
##
## Tucker Lewis Index of factoring reliability =  0.78
## RMSEA index =  0.138  and the 90 % confidence intervals are  0.132 0.145
## BIC =  748.59
## 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 0.98 0.93 0.93 0.88
## Multiple R square of scores with factors          0.94 0.95 0.86 0.86 0.78
## Minimum correlation of possible factor scores     0.88 0.91 0.72 0.72 0.55
##
##  Coefficients and bootstrapped confidence intervals
##                low   PA1 upper   low   PA2 upper   low   PA4 upper   low   PA3
## VERBcomp       0.18  0.26  0.39  0.00  0.05  0.11  0.45  0.59  0.69 -0.01  0.05
## compoundVERBs  0.69  0.80  0.92 -0.06 -0.01  0.05 -0.25 -0.14 -0.03  0.02  0.10
## passives      -0.08  0.03  0.14 -0.05  0.01  0.09 -0.72 -0.59 -0.47  0.12  0.23
## predorder.m   -0.90 -0.79 -0.71 -0.07 -0.03  0.01 -0.10  0.01  0.09 -0.08 -0.01
## predorder.v   -0.64 -0.53 -0.43  0.03  0.10  0.16 -0.05  0.05  0.16  0.07  0.15
## obj           -0.38 -0.29 -0.17 -0.06  0.00  0.05  0.32  0.45  0.55  0.32  0.41
## subj           0.60  0.67  0.75  0.08  0.13  0.17 -0.17 -0.08  0.00 -0.01  0.06
## VERBfrac.m     0.62  0.70  0.79 -0.08 -0.04  0.00  0.32  0.40  0.47 -0.12 -0.07
## NEGcount.m    -0.04  0.03  0.09 -0.14 -0.10 -0.05 -0.22 -0.16 -0.08  0.82  0.90
## NEGcount.v     0.18  0.27  0.33 -0.01  0.04  0.10 -0.24 -0.18 -0.08  0.74  0.81
## NOUNcount.m   -0.96 -0.87 -0.79  0.01  0.04  0.07 -0.20 -0.14 -0.09 -0.24 -0.18
## activity       0.47  0.54  0.62 -0.08 -0.05 -0.01  0.51  0.59  0.66 -0.06 -0.02
## entropy       -0.02  0.03  0.07  0.72  0.77  0.81 -0.02  0.03  0.08  0.07  0.13
## maentropy     -0.21 -0.15 -0.09 -0.03  0.00  0.04 -0.01  0.07  0.14  0.09  0.14
## mamr           0.62  0.71  0.82 -0.08 -0.03  0.01 -0.07  0.01  0.09 -0.09 -0.04
## hapaxes        0.06  0.11  0.16 -0.83 -0.80 -0.76  0.00  0.07  0.13 -0.10 -0.04
## sentcount      0.18  0.24  0.29  0.87  0.90  0.95  0.04  0.09  0.15 -0.29 -0.23
## verbdist      -0.78 -0.70 -0.65 -0.04 -0.01  0.01 -0.48 -0.37 -0.28 -0.23 -0.15
## wordcount     -0.15 -0.12 -0.08  0.91  0.94  0.97 -0.06 -0.03  0.02 -0.02  0.02
##               upper   low   PA5 upper
## VERBcomp       0.12 -0.12 -0.03  0.06
## compoundVERBs  0.18 -0.18 -0.08  0.03
## passives       0.34 -0.22 -0.10  0.01
## predorder.m    0.08 -0.28 -0.15 -0.03
## predorder.v    0.24 -0.16 -0.06  0.04
## obj            0.52 -0.20 -0.11 -0.01
## subj           0.12 -0.30 -0.20 -0.11
## VERBfrac.m    -0.01 -0.09 -0.02  0.03
## NEGcount.m     0.98  0.05  0.12  0.20
## NEGcount.v     0.88  0.04  0.12  0.21
## NOUNcount.m   -0.14 -0.04  0.01  0.07
## activity       0.03 -0.08 -0.03  0.02
## entropy        0.18  0.38  0.44  0.51
## maentropy      0.20  0.62  0.73  0.87
## mamr           0.03 -0.44 -0.31 -0.21
## hapaxes        0.02  0.24  0.31  0.37
## sentcount     -0.18 -0.03  0.03  0.08
## verbdist      -0.05 -0.15 -0.08 -0.02
## wordcount      0.06  0.00  0.04  0.08
##
##  Interfactor correlations and bootstrapped confidence intervals
##           lower estimate upper
## PA1-PA2 -0.0051    0.075 0.181
## PA1-PA4 -0.5679    0.381 0.775
## PA1-PA3 -0.6417   -0.260 0.349
## PA1-PA5 -0.3966   -0.201 0.027
## PA2-PA4 -0.1002    0.113 0.469
## PA2-PA3 -0.0679    0.375 0.573
## PA2-PA5 -0.1258    0.012 0.207
## PA4-PA3 -0.1460    0.075 0.196
## PA4-PA5 -0.5165   -0.283 0.256
## PA3-PA5 -0.3010   -0.036 0.311
```

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: 19 x 2
##    feat          maxload
##    <chr>           <dbl>
##  1 obj             0.449
##  2 predorder.v     0.535
##  3 passives        0.589
##  4 activity        0.589
##  5 VERBcomp        0.590
##  6 subj            0.674
##  7 VERBfrac.m      0.700
##  8 verbdist        0.700
##  9 mamr            0.706
## 10 maentropy       0.729
## 11 entropy         0.767
## 12 predorder.m     0.794
## 13 hapaxes         0.798
## 14 compoundVERBs   0.799
## 15 NEGcount.v      0.807
## 16 NOUNcount.m     0.870
## 17 NEGcount.m      0.896
## 18 sentcount       0.905
## 19 wordcount       0.937
```

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

```
##   predorder.v      passives           obj          subj compoundVERBs
##     0.3411317     0.3454942     0.4739827     0.5190508     0.5618327
##      VERBcomp     maentropy   predorder.m    NEGcount.v          mamr
##     0.5634379     0.5887873     0.6038880     0.6185499     0.7100698
##       hapaxes    NEGcount.m      verbdist   NOUNcount.m     sentcount
##     0.7298809     0.7528934     0.7718060     0.8168441     0.8692217
##       entropy    VERBfrac.m      activity     wordcount
##     0.8702577     0.8902939     0.8902983     0.8915392
```

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

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

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

```
##     wordcount   predorder.m compoundVERBs    NEGcount.m   NOUNcount.m
##      1.037860      1.077369      1.114320      1.124314      1.148401
##     maentropy   predorder.v     sentcount          subj       hapaxes
##      1.184386      1.283173      1.303191      1.308471      1.364657
##      passives    NEGcount.v          mamr      VERBcomp    VERBfrac.m
##      1.376722      1.377108      1.394276      1.419238      1.613803
##       entropy      verbdist      activity           obj
##      1.660458      1.664722      2.003386      2.828878
```

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

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

Feature engineering

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

det(cor(data_engineered_2))
```

```
## [1] 1.575326e-06
```

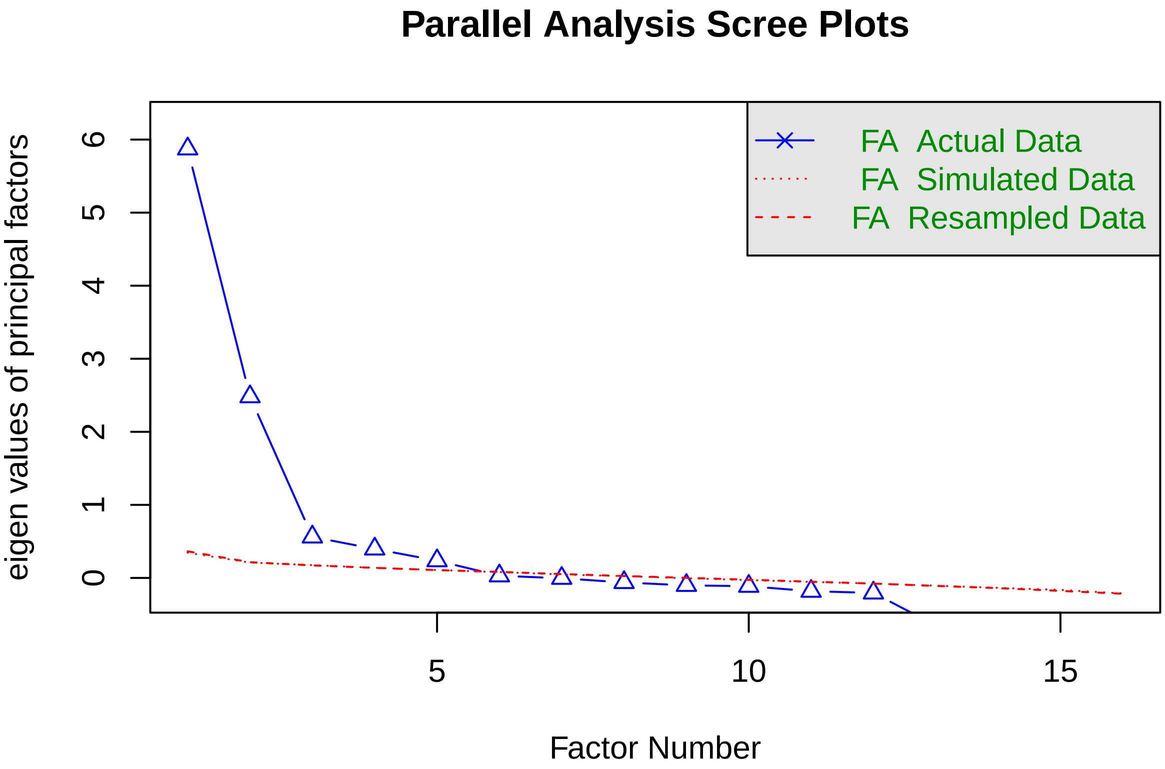
```r
KMO(data_engineered_2)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = data_engineered_2)
## Overall MSA =  0.83
## MSA for each item =
##      VERBcomp compoundVERBs   predorder.m          subj    VERBfrac.m
##          0.84          0.94          0.94          0.93          0.85
##    NEGcount.m    NEGcount.v   NOUNcount.m      activity       entropy
##          0.66          0.64          0.91          0.88          0.72
##     maentropy          mamr       hapaxes     sentcount      verbdist
##          0.62          0.90          0.75          0.72          0.91
##     wordcount
##          0.71
```

Final FA

No. of vectors

```r
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

```
final_collist <- names(data_engineered_2)

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        0.16  0.08  0.59  0.01 -0.01 0.51 0.487 1.2
## compoundVERBs   0.79 -0.05 -0.08  0.01  0.00 0.54 0.464 1.0
## predorder.m    -0.76  0.01  0.02  0.03 -0.12 0.52 0.482 1.1
## subj            0.75  0.11 -0.16  0.01 -0.14 0.54 0.461 1.2
## VERBfrac.m     0.59 -0.06  0.44 -0.06 -0.03 0.90 0.098 1.9
## NEGcount.m    -0.11 -0.06  0.04  0.92 -0.01 0.85 0.150 1.0
## NEGcount.v     0.16  0.07 -0.03  0.79  0.01 0.66 0.339 1.1
## NOUNcount.m   -0.88  0.07 -0.09 -0.10 -0.03 0.84 0.165 1.1
## activity       0.38 -0.04  0.66  0.01 -0.06 0.91 0.092 1.6
## entropy        0.10  0.74 -0.05  0.03  0.46 0.89 0.110 1.7
## maentropy     -0.06 -0.05 -0.03  0.00  0.82 0.70 0.301 1.0
## mamr           0.73 -0.05 -0.01 -0.06 -0.25 0.71 0.291 1.2
## hapaxes        0.15 -0.83 -0.01 -0.10  0.31 0.71 0.288 1.4
## sentcount      0.21  0.85  0.11 -0.16  0.00 0.83 0.172 1.2
## verbdist      -0.69 -0.01 -0.29 -0.07 -0.10 0.75 0.246 1.4
## wordcount     -0.14  0.94  0.01  0.03  0.03 0.89 0.107 1.0
##
##                        PA1  PA2  PA5  PA3  PA4
## SS loadings           4.63 2.89 1.56 1.53 1.14
## Proportion Var        0.29 0.18 0.10 0.10 0.07
## Cumulative Var        0.29 0.47 0.57 0.66 0.73
## 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 0.15  0.61 -0.16 -0.29
## PA2  0.15 1.00  0.06  0.31  0.14
## PA5  0.61 0.06  1.00 -0.17 -0.16
## PA3 -0.16 0.31 -0.17  1.00  0.27
## PA4 -0.29 0.14 -0.16  0.27  1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 5 factors are sufficient.
##
## df null model =  120  with the objective function =  13.36 with Chi Square =  9965.12
## df of  the model are 50  and the objective function was  0.87
##
## 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  68.63  with prob <  0.041
## The total n.obs was  753  with Likelihood Chi Square =  642.69  with prob <  7.1e-104
##
## Tucker Lewis Index of factoring reliability =  0.855
## RMSEA index =  0.125  and the 90 % confidence intervals are  0.117 0.134
## BIC =  311.49
## 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 0.98 0.94 0.94 0.91
## Multiple R square of scores with factors          0.94 0.95 0.89 0.89 0.83
## Minimum correlation of possible factor scores     0.88 0.91 0.77 0.78 0.65
##
##  Coefficients and bootstrapped confidence intervals
##                 low   PA1 upper   low   PA2 upper   low   PA5 upper   low   PA3
## VERBcomp       0.05  0.16  0.30  0.03  0.08  0.14  0.42  0.59  0.80 -0.04  0.01
## compoundVERBs  0.70  0.79  0.86 -0.09 -0.05  0.01 -0.17 -0.08  0.01 -0.04  0.01
## predorder.m   -0.90 -0.76 -0.62 -0.04  0.01  0.05 -0.12  0.02  0.12 -0.05  0.03
## subj           0.66  0.75  0.83  0.06  0.11  0.15 -0.28 -0.16 -0.04 -0.04  0.01
## VERBfrac.m     0.52  0.59  0.68 -0.09 -0.06 -0.03  0.31  0.44  0.60 -0.10 -0.06
## NEGcount.m    -0.16 -0.11 -0.07 -0.10 -0.06 -0.03  0.00  0.04  0.09  0.86  0.92
## NEGcount.v     0.11  0.16  0.22  0.03  0.07  0.11 -0.09 -0.03  0.04  0.71  0.79
## NOUNcount.m   -0.99 -0.88 -0.76  0.03  0.07  0.10 -0.20 -0.09 -0.02 -0.15 -0.10
## activity       0.31  0.38  0.48 -0.07 -0.04 -0.01  0.47  0.66  0.88 -0.02  0.01
## entropy        0.04  0.10  0.15  0.71  0.74  0.78 -0.11 -0.05  0.00  0.00  0.03
## maentropy     -0.10 -0.06 -0.01 -0.08 -0.05 -0.02 -0.11 -0.03  0.02 -0.03  0.00
## mamr           0.65  0.73  0.82 -0.10 -0.05  0.00 -0.13 -0.01  0.09 -0.11 -0.06
## hapaxes        0.08  0.15  0.21 -0.86 -0.83 -0.80 -0.08 -0.01  0.06 -0.14 -0.10
## sentcount      0.14  0.21  0.28  0.81  0.85  0.89  0.04  0.11  0.19 -0.19 -0.16
## verbdist      -0.78 -0.69 -0.61 -0.03 -0.01  0.03 -0.47 -0.29 -0.15 -0.13 -0.07
## wordcount     -0.18 -0.14 -0.10  0.92  0.94  0.97 -0.03  0.01  0.05  0.00  0.03
##               upper   low   PA4 upper
## VERBcomp       0.06 -0.08 -0.01  0.05
## compoundVERBs  0.07 -0.06  0.00  0.07
## predorder.m    0.13 -0.18 -0.12 -0.05
## subj           0.07 -0.21 -0.14 -0.07
## VERBfrac.m    -0.02 -0.07 -0.03  0.01
## NEGcount.m     1.01 -0.04 -0.01  0.03
## NEGcount.v     0.87 -0.03  0.01  0.06
## NOUNcount.m   -0.05 -0.07 -0.03  0.00
## activity       0.04 -0.10 -0.06 -0.02
## entropy        0.07  0.40  0.46  0.52
## maentropy      0.04  0.73  0.82  0.91
## mamr           0.00 -0.30 -0.25 -0.18
## hapaxes       -0.06  0.25  0.31  0.36
## sentcount     -0.13 -0.03  0.00  0.04
## verbdist      -0.01 -0.15 -0.10 -0.05
## wordcount      0.05  0.00  0.03  0.06
##
##  Interfactor correlations and bootstrapped confidence intervals
##           lower estimate upper
## PA1-PA2  0.0086    0.147  0.30
## PA1-PA5 -0.6228    0.610  0.94
## PA1-PA3 -0.6851   -0.163  0.82
## PA1-PA4 -0.6300   -0.289  0.27
## PA2-PA5 -0.0380    0.055  0.46
## PA2-PA3 -0.0775    0.313  0.41
## PA2-PA4 -0.0707    0.144  0.28
## PA5-PA3 -0.4339   -0.173  0.30
## PA5-PA4 -0.4064   -0.163  0.43
## PA3-PA4 -0.3589    0.271  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        0.595
##  2 VERBfrac.m      0.595
##  3 activity        0.663
##  4 verbdist        0.694
##  5 mamr            0.733
##  6 entropy         0.741
##  7 subj            0.746
##  8 predorder.m     0.756
##  9 compoundVERBs   0.786
## 10 NEGcount.v      0.792
## 11 maentropy       0.815
## 12 hapaxes         0.826
## 13 sentcount       0.854
## 14 NOUNcount.m     0.885
## 15 NEGcount.m      0.924
## 16 wordcount       0.940
```

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

```
##      VERBcomp   predorder.m compoundVERBs          subj    NEGcount.v 
##     0.5127379     0.5184379     0.5355550     0.5388283     0.6612655 
##     maentropy          mamr       hapaxes      verbdist     sentcount 
##     0.6992318     0.7090855     0.7124561     0.7542810     0.8278966 
##   NOUNcount.m    NEGcount.m       entropy     wordcount    VERBfrac.m 
##     0.8351383     0.8496804     0.8902622     0.8931153     0.9024173 
##      activity 
##     0.9082612
```

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

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

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

```
##     maentropy compoundVERBs    NEGcount.m     wordcount   predorder.m 
##      1.021058      1.027312      1.044492      1.047851      1.059200 
##   NOUNcount.m    NEGcount.v      VERBcomp          subj     sentcount 
##      1.062351      1.105014      1.186239      1.205567      1.224178 
##          mamr       hapaxes      verbdist      activity       entropy 
##      1.246067      1.391271      1.405896      1.613767      1.737988 
##    VERBfrac.m 
##      1.899946
```

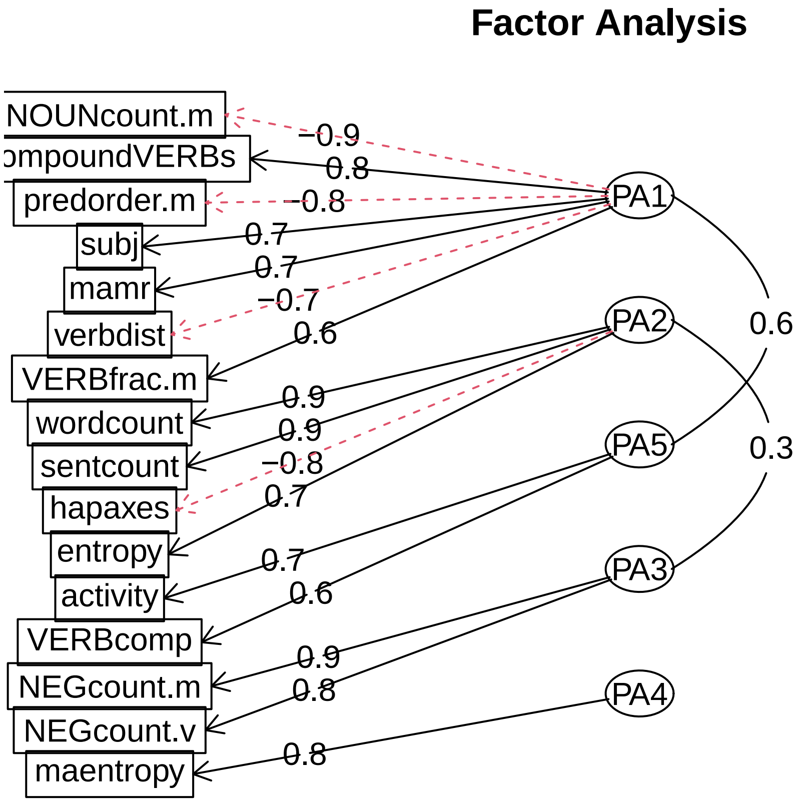
```
fa_res$complexity[fa_res$complexity > 2] %>% names()
```

```
## 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       0.160         0.595              
## compoundVERBs  0.786                            
## predorder.m   -0.756                      -0.124
## subj           0.746  0.108 -0.156        -0.140
## VERBfrac.m     0.595         0.441              
## NEGcount.m    -0.114                0.924       
## NEGcount.v     0.165                0.792       
## NOUNcount.m   -0.885               -0.100       
## activity       0.377         0.663              
## entropy               0.741                0.463
## maentropy                                  0.815
## mamr           0.733                      -0.245
## hapaxes        0.154 -0.826                0.312
## sentcount      0.206  0.854  0.106 -0.160       
## verbdist      -0.694        -0.287        -0.102
## wordcount     -0.139  0.940                     
## 
##                  PA1   PA2   PA5   PA3   PA4
## SS loadings    4.206 2.888 1.128 1.542 1.089
## Proportion Var 0.263 0.181 0.070 0.096 0.068
## Cumulative Var 0.263 0.443 0.514 0.610 0.678
```

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

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

  load_df_filtered <- load_df %>%
    mutate(abs_l = abs(loading)) %>%
    mutate(strng = case_when(
      abs_l > 0.70 ~ "*****",
      abs_l <= 0.70 & abs_l > 0.63 ~ "**** ",
      abs_l <= 0.63 & abs_l > 0.55 ~ "***  ",
      abs_l <= 0.55 & abs_l > 0.45 ~ "**   ",
      abs_l <= 0.45 & abs_l > 0.32 ~ "*    ",
      .default = ""
    )) %>%
    arrange(-abs_l) %>%
    filter(abs_l > 0.1)

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

  cat("\n")
}
```

```
##
## ----- PA1 -----
##               loading abs_l strng
## NOUNcount.m    -0.885 0.885 *****
## compoundVERBs   0.786 0.786 *****
## predorder.m    -0.756 0.756 *****
## subj            0.746 0.746 *****
## mamr            0.733 0.733 *****
## verbdist       -0.694 0.694 ****
## VERBfrac.m      0.595 0.595 ***
## activity        0.377 0.377 *
## sentcount       0.206 0.206
## NEGcount.v      0.165 0.165
## VERBcomp        0.160 0.160
## hapaxes         0.154 0.154
## wordcount      -0.139 0.139
## NEGcount.m     -0.114 0.114
##
##
## ----- PA2 -----
##           loading abs_l strng
## wordcount   0.940 0.940 *****
## sentcount   0.854 0.854 *****
## hapaxes    -0.826 0.826 *****
## entropy     0.741 0.741 *****
## subj        0.108 0.108
##
##
```

```
## ----- PA5 -----
##            loading abs_l strng
## activity     0.663 0.663 ****
## VERBcomp     0.595 0.595 ***
## VERBfrac.m   0.441 0.441 *
## verbdist    -0.287 0.287
## subj        -0.156 0.156
## sentcount    0.106 0.106
##
##
## ----- PA3 -----
##             loading abs_l strng
## NEGcount.m    0.924 0.924 *****
## NEGcount.v    0.792 0.792 *****
## sentcount    -0.160 0.160
## NOUNcount.m  -0.100 0.100
##
##
## ----- PA4 -----
##             loading abs_l strng
## maentropy     0.815 0.815 *****
## entropy       0.463 0.463 **
## hapaxes       0.312 0.312
## mamr         -0.245 0.245
## subj         -0.140 0.140
## predorder.m  -0.124 0.124
## verbdist     -0.102 0.102
```

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~~ not anymore
  - ~~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)
```

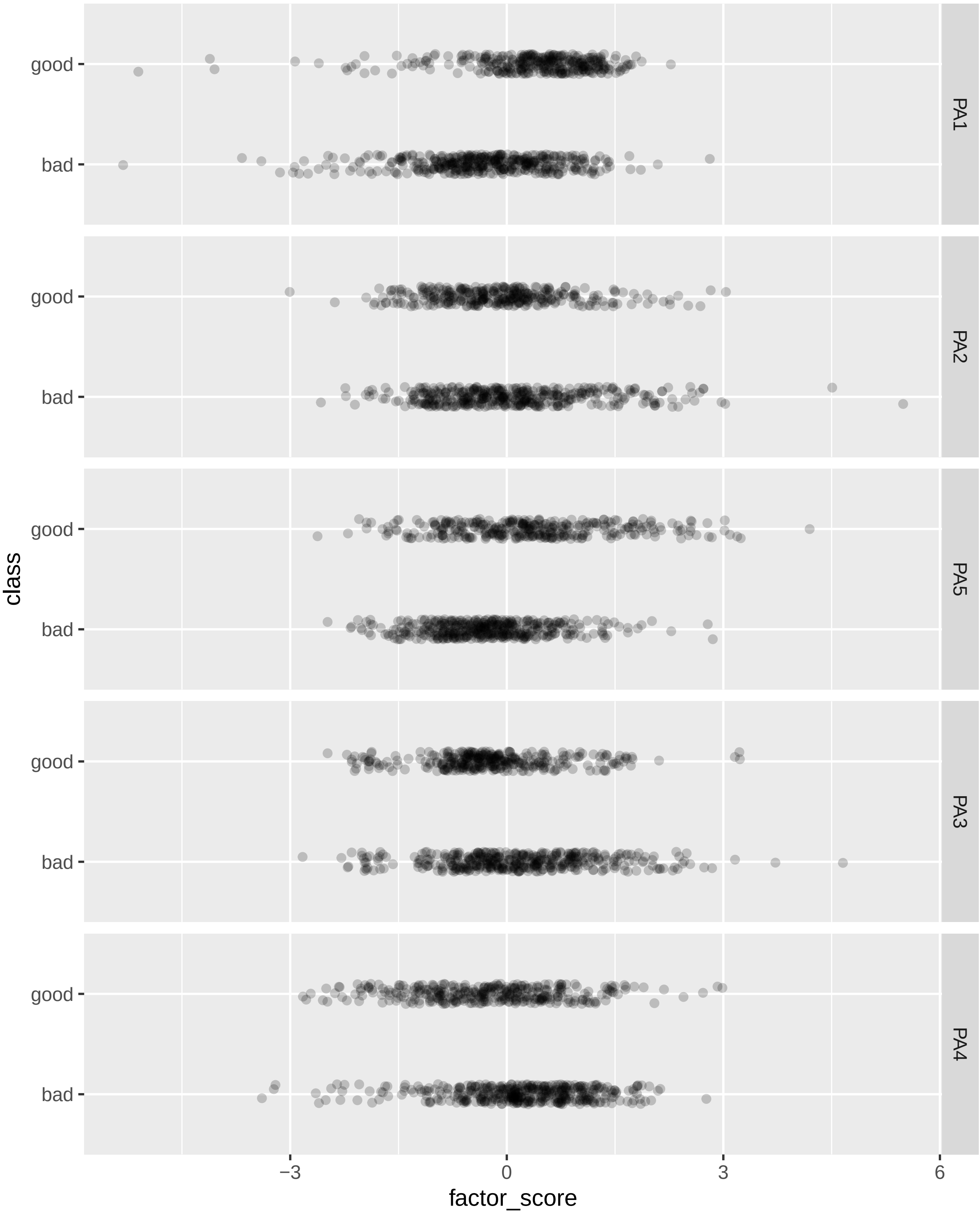
```
##      VERBcomp compoundVERBs   predorder.m          subj    VERBfrac.m
##         0.487         0.464         0.482         0.461         0.098
##    NEGcount.m    NEGcount.v   NOUNcount.m      activity       entropy
##         0.150         0.339         0.165         0.092         0.110
##     maentropy          mamr       hapaxes     sentcount      verbdist
##         0.301         0.291         0.288         0.172         0.246
##     wordcount
##         0.107
```

Distributions over factors

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

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

  p_val <- numeric()
  epsilon2 <- numeric()
  min_p_values <- numeric()
  for (f in factors) {
    data <- data_factors_long %>% filter(factor == f)

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

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

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

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

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

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

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

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

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

data_factors_long %>%
  group_by(factor) %>%
  summarize(shapiro = shapiro.test(factor_score)$p.value)
```

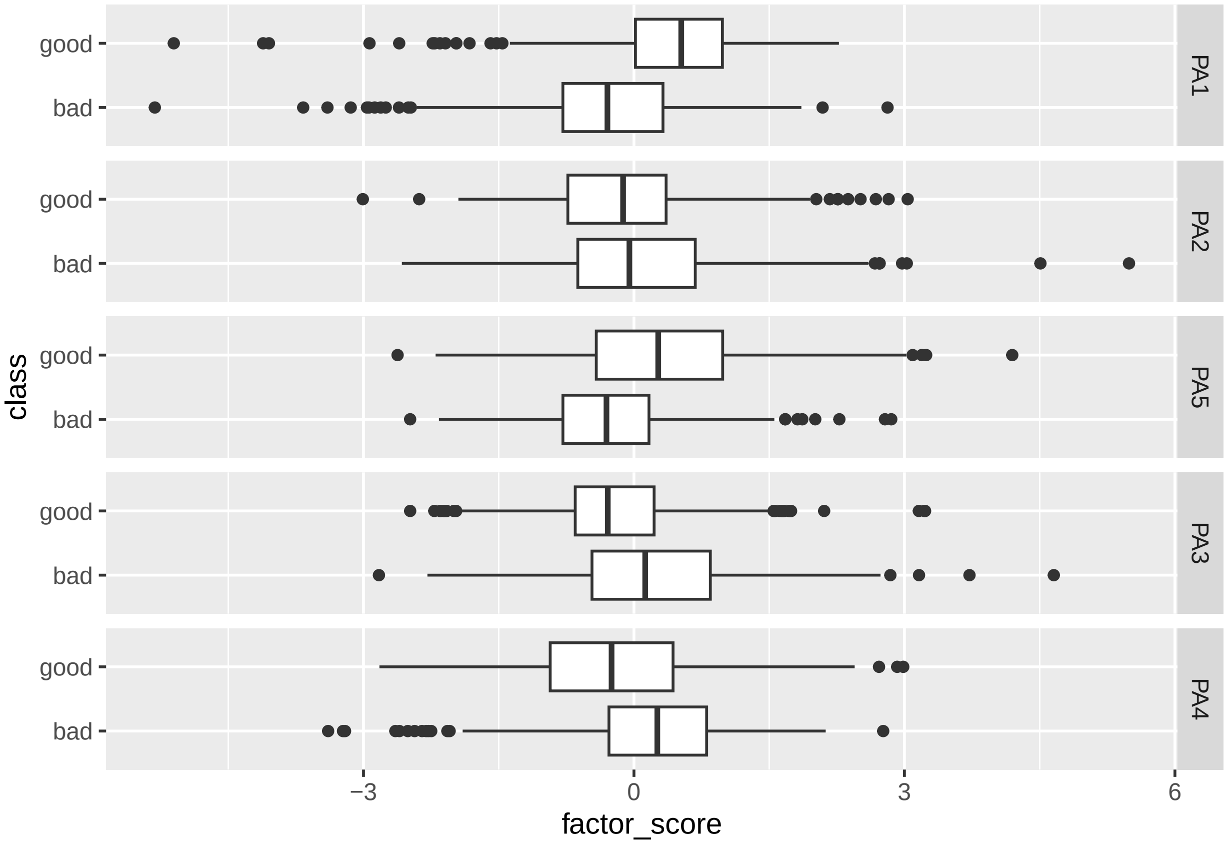
```
## # A tibble: 5 x 2
##   factor  shapiro
##   <fct>     <dbl>
## 1 PA1    2.42e-15
## 2 PA2    2.99e-11
## 3 PA5    2.22e- 9
## 4 PA3    9.41e- 9
## 5 PA4    4.57e- 5
```

```
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")
```



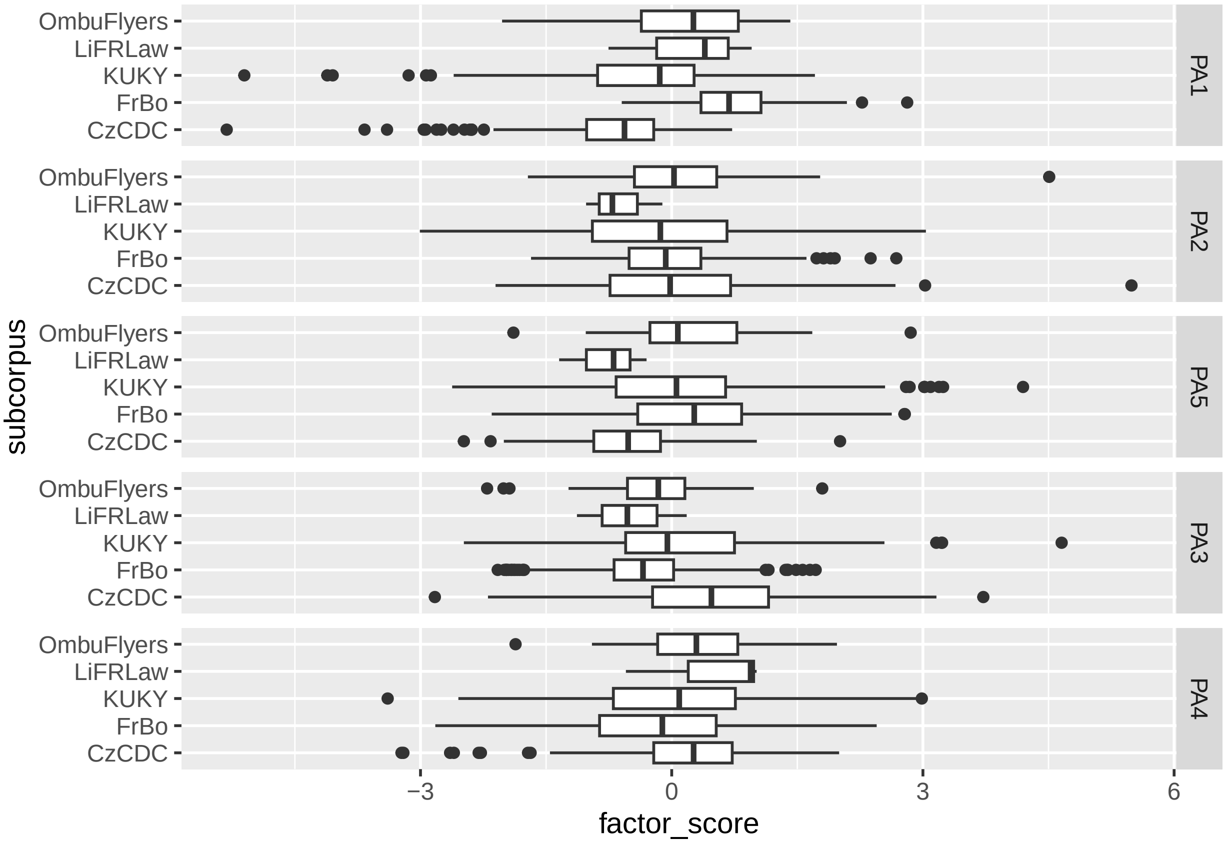
```
##
## Test for the significance of differences in class over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 121.9287, df = 1, p-value = 0
##
##
##                              Comparison of x by group
##                                    (Bonferroni)
## Col Mean-|
## Row Mean |        bad
## ---------+-----------
##     good |  -11.04213
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.162
##
## Test for the significance of differences in class over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 3.9267, df = 1, p-value = 0.05
```

```
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |        bad
## ---------+-----------
##     good |   1.981593
##          |    0.0475*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.00522
##
## Test for the significance of differences in class over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 67.2231, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |        bad
## ---------+-----------
##     good |  -8.198970
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0894
##
## Test for the significance of differences in class over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 31.3255, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |        bad
## ---------+-----------
##     good |   5.596919
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0417
##
```

```
## Test for the significance of differences in class over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 47.3983, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |        bad
## ---------+-----------
##     good |   6.884643
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.063
##
##   factor kruskal_p epsilon2
## 1    PA1  2.39e-28  0.16200
## 2    PA2  4.75e-02  0.00522
## 3    PA5  2.42e-16  0.08940
## 4    PA3  2.18e-08  0.04170
## 5    PA4  5.79e-12  0.06300
##
## 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

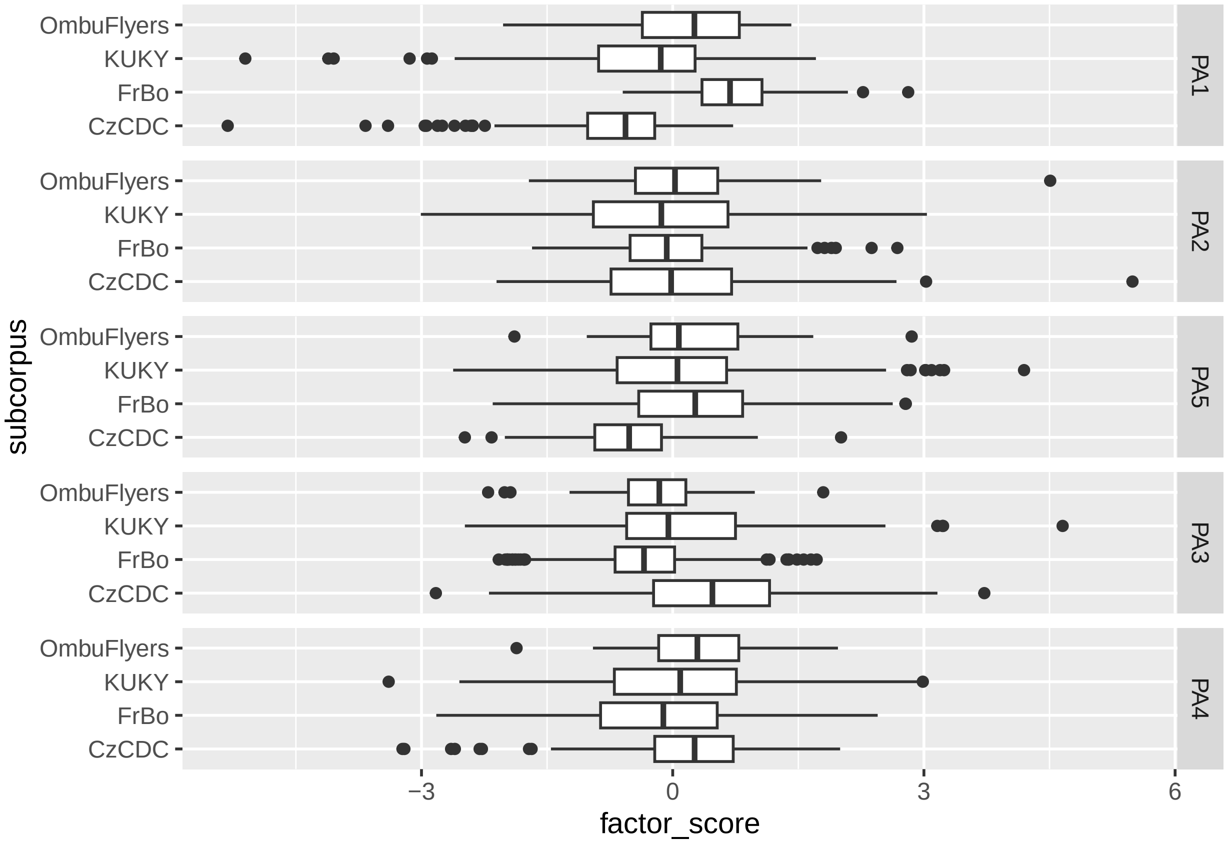
```
analyze_distributions(data_factors_long, "subcorpus")
```



```
##
## Test for the significance of differences in subcorpus over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 366.863, df = 4, p-value = 0
##
##
##                              Comparison of x by group
##                                    (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY    LiFRLaw
## ---------+--------------------------------------------
##     FrBo |  -18.06472
##          |    0.0000*
##          |
##     KUKY |  -4.318421   12.92974
##          |    0.0002*    0.0000*
##          |
##  LiFRLaw |  -1.713558   1.067093  -0.974197
##          |     0.8661     1.0000     1.0000
##          |
## OmbuFlye |  -5.613026   3.641762  -3.154508   0.011969
##          |    0.0000*    0.0027*    0.0161*     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.488
##
## Test for the significance of differences in subcorpus over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 5.6768, df = 4, p-value = 0.22
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY    LiFRLaw
## ---------+--------------------------------------------
##     FrBo |   0.555047
##          |     1.0000
##          |
##     KUKY |   1.677136   1.277711
##          |     0.9352     1.0000
##          |
##  LiFRLaw |   1.383570   1.301060   1.095985
##          |     1.0000     1.0000     1.0000
##          |
## OmbuFlye |  -0.584188  -0.887273  -1.520699  -1.513090
##          |     1.0000     1.0000     1.0000     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.00755
##
## Test for the significance of differences in subcorpus over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 111.5455, df = 4, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY    LiFRLaw
## ---------+--------------------------------------------
##     FrBo |  -10.13360
##          |    0.0000*
##          |
##     KUKY |  -6.885610   2.412725
##          |    0.0000*     0.1583
##          |
##  LiFRLaw |   0.509734   2.072806   1.686637
##          |     1.0000     0.3819     0.9167
##          |
## OmbuFlye |  -5.020268   0.124985  -1.126239  -1.969418
##          |    0.0000*     1.0000     1.0000     0.4891
##
## 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 = 96.1298, df = 4, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY    LiFRLaw
## ---------+--------------------------------------------
##     FrBo |   9.694520
##          |    0.0000*
##          |
##     KUKY |   4.667193  -4.390526
##          |    0.0000*    0.0001*
##          |
##  LiFRLaw |   1.883974   0.393844   1.084879
##          |     0.5957     1.0000     1.0000
##          |
## OmbuFlye |   3.666503  -1.283929   1.025313  -0.749181
##          |    0.0025*     1.0000     1.0000     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.128
##
## Test for the significance of differences in subcorpus over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 24.5474, df = 4, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY    LiFRLaw
## ---------+--------------------------------------------
##     FrBo |   4.443569
##          |    0.0001*
##          |
##     KUKY |   1.957020  -2.210067
##          |     0.5035     0.2710
##          |
##  LiFRLaw |  -0.547145  -1.233261  -0.881396
##          |     1.0000     1.0000     1.0000
##          |
## OmbuFlye |  -0.553727  -2.878101  -1.647371   0.367765
##          |     1.0000    0.0400*     0.9948     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0326
##
##   factor kruskal_p epsilon2
## 1    PA1  4.00e-78  0.48800
## 2    PA2  2.25e-01  0.00755
## 3    PA5  3.41e-23  0.14800
## 4    PA3  6.55e-20  0.12800
## 5    PA4  6.20e-05  0.03260
##
## 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
```

subcorpus wo/ LiFRLaw

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



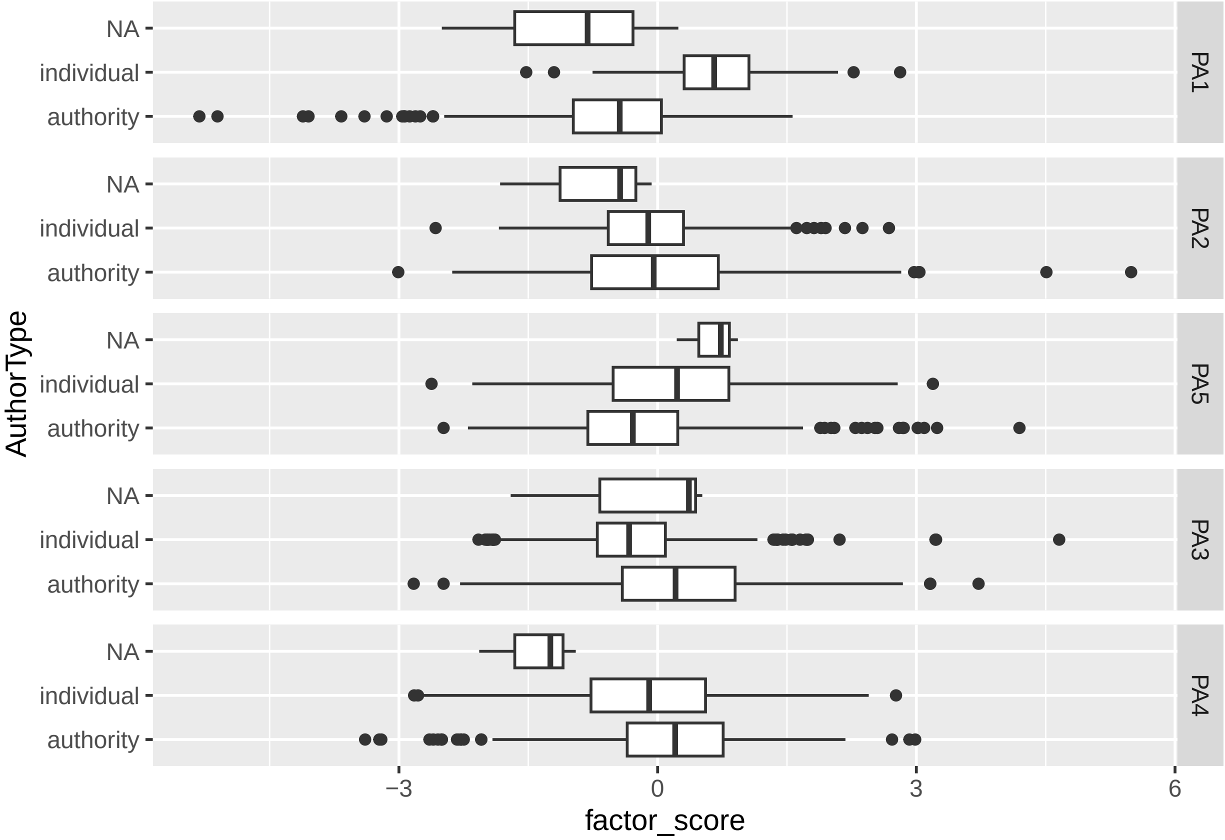
```
##
## Test for the significance of differences in subcorpus over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 366.7061, df = 3, p-value = 0
##
##
##                              Comparison of x by group
##                                    (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY
## ---------+---------------------------------
##     FrBo |  -18.06396
##          |    0.0000*
##          |
##     KUKY |  -4.320184   12.92709
##          |    0.0001*    0.0000*
##          |
## OmbuFlye |  -5.610052   3.644413  -3.150565
##          |    0.0000*    0.0016*    0.0098*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.49
##
```

```
## Test for the significance of differences in subcorpus over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 3.984, df = 3, p-value = 0.26
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY
## ---------+---------------------------------
##     FrBo |   0.566603
##          |     1.0000
##          |
##     KUKY |   1.674476   1.263559
##          |     0.5642     1.0000
##          |
## OmbuFlye |  -0.578350  -0.887300  -1.513408
##          |     1.0000     1.0000     0.7811
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.00532
##
## Test for the significance of differences in subcorpus over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 109.2883, df = 3, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY
## ---------+---------------------------------
##     FrBo |  -10.13874
##          |    0.0000*
##          |
##     KUKY |  -6.891583   2.411255
##          |    0.0000*     0.0954
##          |
## OmbuFlye |  -5.019324   0.128623  -1.121953
##          |    0.0000*     1.0000     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.146
##
## Test for the significance of differences in subcorpus over PA3 :
##
```

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 95.1198, df = 3, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY
## ---------+---------------------------------
##     FrBo |   9.695573
##          |    0.0000*
##          |
##     KUKY |   4.665674  -4.393200
##          |    0.0000*    0.0001*
##          |
## OmbuFlye |   3.665920  -1.285074   1.025586
##          |    0.0015*     1.0000     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.127
##
## Test for the significance of differences in subcorpus over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 23.7598, df = 3, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC       FrBo       KUKY
## ---------+---------------------------------
##     FrBo |   4.449040
##          |    0.0001*
##          |
##     KUKY |   1.961852  -2.210161
##          |     0.2987     0.1626
##          |
## OmbuFlye |  -0.550139  -2.877270  -1.646516
##          |     1.0000    0.0241*     0.5979
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0317
##
##   factor kruskal_p epsilon2
## 1    PA1  3.60e-79  0.49000
## 2    PA2  2.63e-01  0.00532
## 3    PA5  1.56e-23  0.14600
## 4    PA3  1.74e-20  0.12700
## 5    PA4  2.80e-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 PA4
```

AuthorType

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

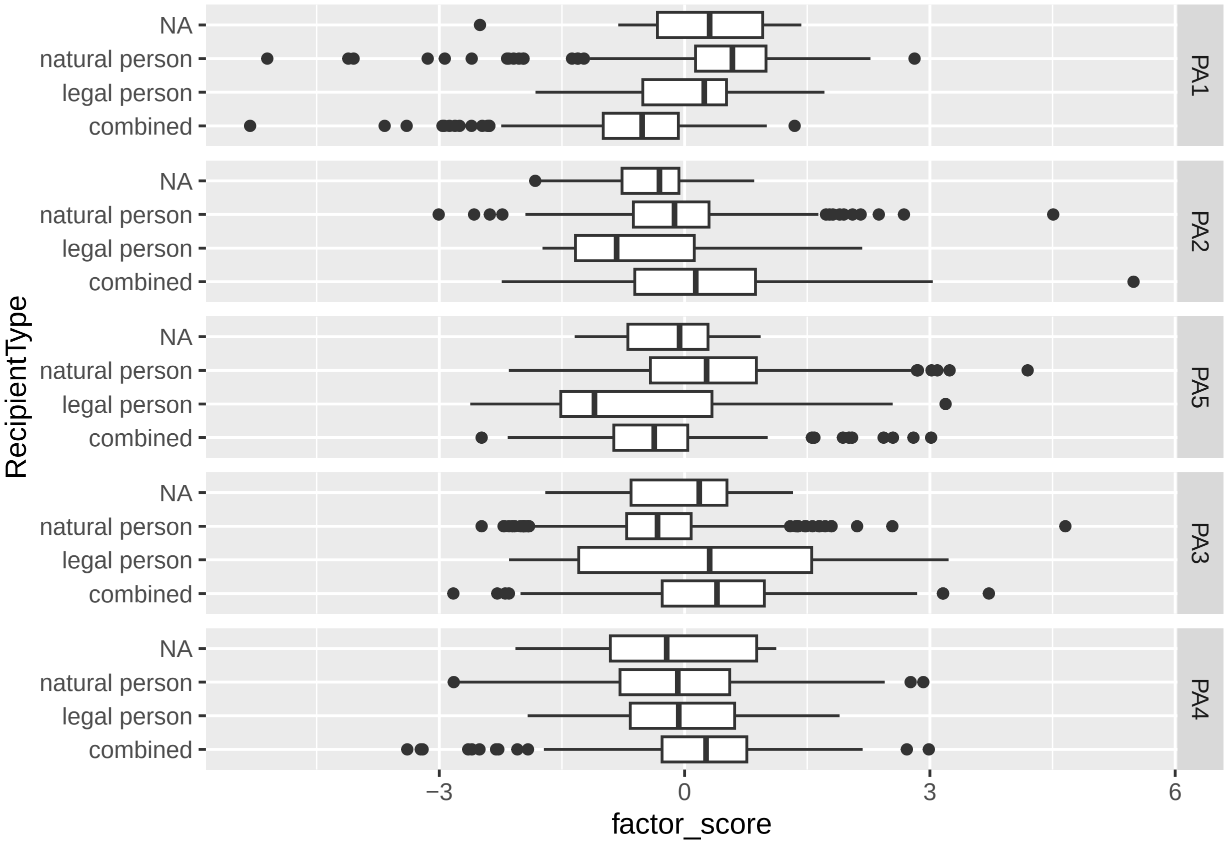


```
##
## Test for the significance of differences in AuthorType over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 340.9066, df = 1, p-value = 0
##
##
##                              Comparison of x by group
##                                    (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## ---------+-----------
## individu |  -18.46365
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.453
##
## Test for the significance of differences in AuthorType over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 1.2713, df = 1, p-value = 0.26
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## ---------+-----------
## individu |   1.127532
##          |     0.2595
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.00169
##
## Test for the significance of differences in AuthorType over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 41.6472, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## ---------+-----------
## individu |  -6.453466
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0554
##
## Test for the significance of differences in AuthorType over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 57.8083, df = 1, p-value = 0
##
##
```

```
##                             Comparison of x by group
##                                   (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## ---------+-----------
## individu |   7.603179
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0769
##
## Test for the significance of differences in AuthorType over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 19.6252, df = 1, p-value = 0
##
##
##                             Comparison of x by group
##                                   (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## ---------+-----------
## individu |   4.430037
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0261
##
##   factor kruskal_p epsilon2
## 1    PA1  4.05e-76  0.45300
## 2    PA2  2.60e-01  0.00169
## 3    PA5  1.09e-10  0.05540
## 4    PA3  2.89e-14  0.07690
## 5    PA4  9.42e-06  0.02610
##
## 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

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



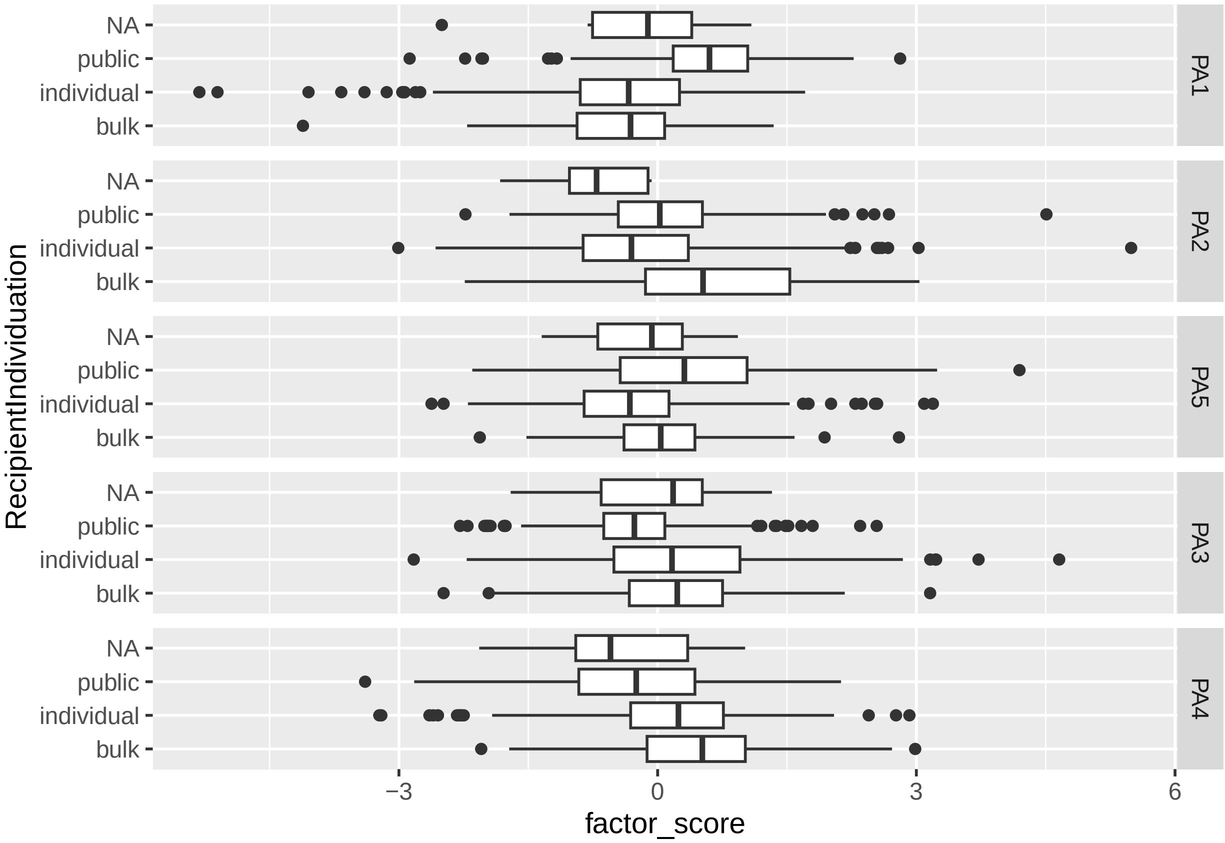
```
##
## Test for the significance of differences in RecipientType over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 271.8125, df = 2, p-value = 0
##
##
##                           Comparison of x by group
##                                 (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## ---------+----------------------
## legal pe |  -3.491563
##          |    0.0014*
##          |
## natural  |  -16.48578  -2.290617
##          |    0.0000*     0.0660
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.361
##
## Test for the significance of differences in RecipientType over PA2 :
##
##   Kruskal-Wallis rank sum test
```

```
##
## data: x and group
## Kruskal-Wallis chi-squared = 21.206, df = 2, p-value = 0
##
##
##                             Comparison of x by group
##                                   (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## ---------+----------------------
## legal pe |   3.808299
##          |    0.0004*
##          |
## natural  |   3.302548  -2.679236
##          |    0.0029*    0.0221*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0282
##
## Test for the significance of differences in RecipientType over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 93.0143, df = 2, p-value = 0
##
##
##                             Comparison of x by group
##                                   (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## ---------+----------------------
## legal pe |   0.193176
##          |     1.0000
##          |
## natural  |  -9.406223  -3.512859
##          |    0.0000*    0.0013*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.124
##
## Test for the significance of differences in RecipientType over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 99.3289, df = 2, p-value = 0
##
##
##                             Comparison of x by group
##                                   (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## ---------+----------------------
## legal pe |   1.274923
##          |     0.6070
##          |
## natural  |   9.938824   2.218805
##          |    0.0000*     0.0795
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.132
##
## Test for the significance of differences in RecipientType over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 21.8926, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## ---------+----------------------
## legal pe |   1.464990
##          |     0.4288
##          |
## natural  |   4.647620   0.160578
##          |    0.0000*     1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0291
##
##   factor kruskal_p epsilon2
## 1    PA1  9.48e-60   0.3610
## 2    PA2  2.48e-05   0.0282
## 3    PA5  6.34e-21   0.1240
## 4    PA3  2.70e-22   0.1320
## 5    PA4  1.76e-05   0.0291
##
## p < 5e-2 found in: PA1 PA2 PA5 PA3 PA4
## p < 1e-2 found in: PA1 PA2 PA5 PA3 PA4
## p < 1e-3 found in: PA1 PA2 PA5 PA3 PA4
## p < 1e-4 found in: PA1 PA5 PA3 PA4
```

court decisions often with `RecipientType = combined`.

RecipientIndividuation

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

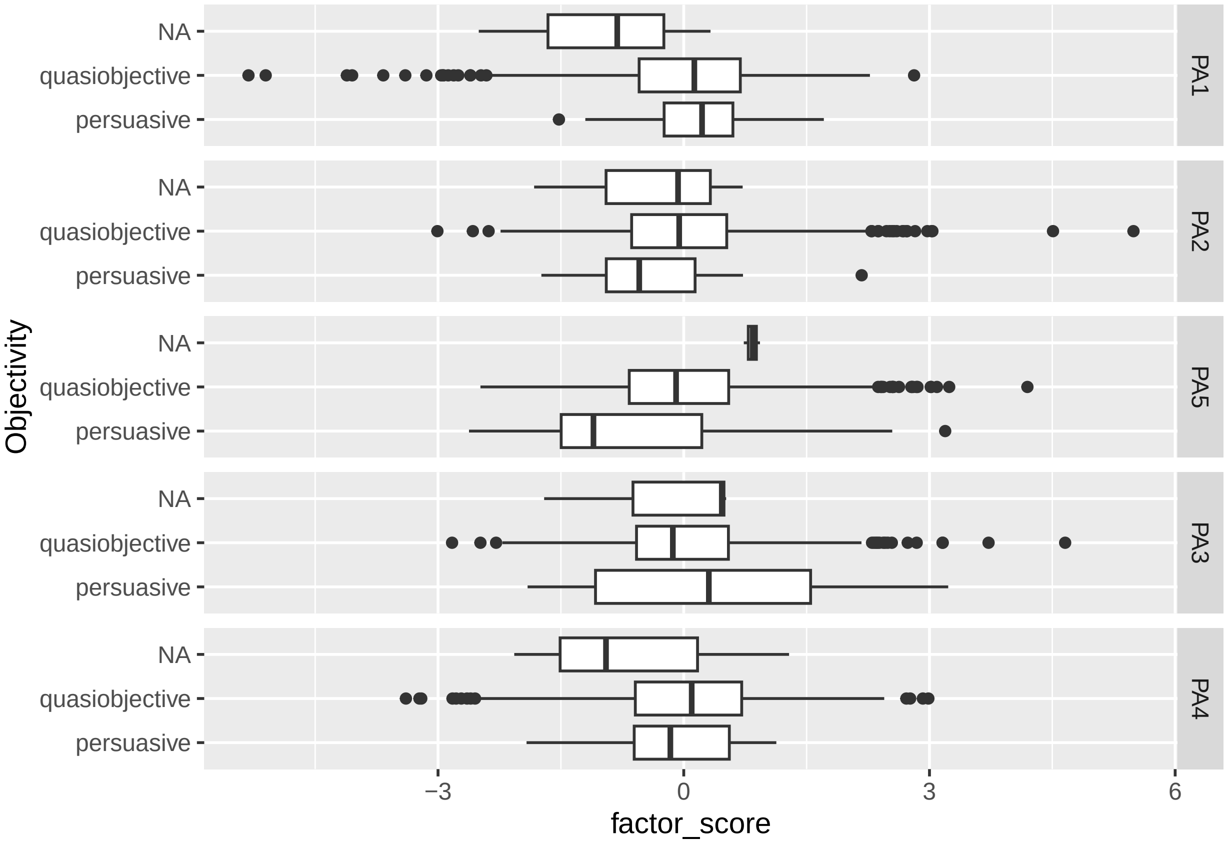


```
##
## Test for the significance of differences in RecipientIndividuation over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 204.8087, df = 2, p-value = 0
##
##
##                              Comparison of x by group
##                                    (Bonferroni)
## Col Mean-|
## Row Mean |       bulk   individu
## ---------+----------------------
## individu |  -0.708707
##          |     1.0000
##          |
##   public |  -8.563337  -13.53797
##          |    0.0000*    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.272
##
## Test for the significance of differences in RecipientIndividuation over PA2 :
##
##   Kruskal-Wallis rank sum test
```

```
##
## data: x and group
## Kruskal-Wallis chi-squared = 36.9687, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |       bulk   individu
## ---------+----------------------
## individu |   5.608745
##          |    0.0000*
##          |
##   public |   3.344736  -3.809224
##          |    0.0025*    0.0004*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0492
##
## Test for the significance of differences in RecipientIndividuation over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 74.3427, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |       bulk   individu
## ---------+----------------------
## individu |   2.974126
##          |    0.0088*
##          |
##   public |  -2.047427  -8.600186
##          |     0.1218    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0989
##
## Test for the significance of differences in RecipientIndividuation over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 42.7107, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |       bulk   individu
## ---------+----------------------
## individu |   0.645876
##          |     1.0000
##          |
##   public |   4.164581   6.069974
##          |    0.0001*    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0568
##
## Test for the significance of differences in RecipientIndividuation over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 48.0777, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |       bulk   individu
## ---------+----------------------
## individu |   1.577959
##          |     0.3437
##          |
##   public |   5.076752   6.050584
##          |    0.0000*    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0639
##
##   factor kruskal_p epsilon2
## 1    PA1  3.36e-45   0.2720
## 2    PA2  9.38e-09   0.0492
## 3    PA5  7.19e-17   0.0989
## 4    PA3  5.31e-10   0.0568
## 5    PA4  3.63e-11   0.0639
##
## p < 5e-2 found in: PA1 PA2 PA5 PA3 PA4
## p < 1e-2 found in: PA1 PA2 PA5 PA3 PA4
## p < 1e-3 found in: PA1 PA2 PA5 PA3 PA4
## p < 1e-4 found in: PA1 PA2 PA5 PA3 PA4
```

Objectivity

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



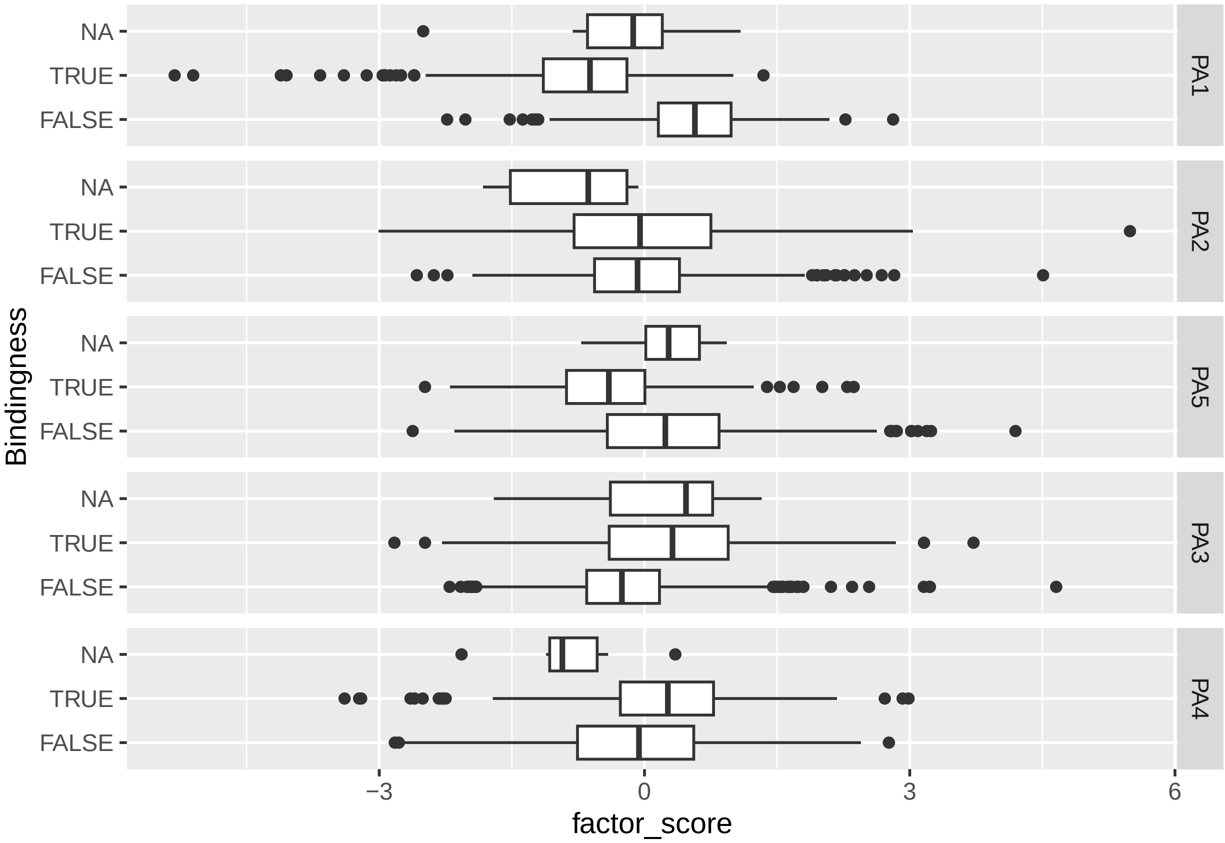
```
##
## Test for the significance of differences in Objectivity over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.3457, df = 1, p-value = 0.56
##
##
##                              Comparison of x by group
##                                    (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## ---------+-----------
## quasiobj |   0.587952
##          |     0.5566
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.00046
##
## Test for the significance of differences in Objectivity over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 5.1139, df = 1, p-value = 0.02
```

```
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## ---------+-----------
## quasiobj |  -2.261396
##          |    0.0237*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0068
##
## Test for the significance of differences in Objectivity over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 5.4616, df = 1, p-value = 0.02
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## ---------+-----------
## quasiobj |  -2.336998
##          |    0.0194*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.00726
##
## Test for the significance of differences in Objectivity over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.6164, df = 1, p-value = 0.43
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## ---------+-----------
## quasiobj |   0.785129
##          |     0.4324
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.00082
##
```

```
## Test for the significance of differences in Objectivity over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.8539, df = 1, p-value = 0.36
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## ---------+-----------
## quasiobj |  -0.924072
##          |     0.3554
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.00114
##
##   factor kruskal_p epsilon2
## 1    PA1    0.5566  0.00046
## 2    PA2    0.0237  0.00680
## 3    PA5    0.0194  0.00726
## 4    PA3    0.4324  0.00082
## 5    PA4    0.3554  0.00114
##
## 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 = 349.4445, df = 1, p-value = 0
##
##
##                              Comparison of x by group
##                                    (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## ---------+-----------
##     TRUE |   18.69343
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.465
##
## Test for the significance of differences in Bindingness over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.5482, df = 1, p-value = 0.46
```

```
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## ---------+-----------
##     TRUE |  -0.740375
##          |     0.4591
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.000729
##
## Test for the significance of differences in Bindingness over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 97.6022, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## ---------+-----------
##     TRUE |   9.879380
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.13
##
## Test for the significance of differences in Bindingness over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 49.5731, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## ---------+-----------
##     TRUE |  -7.040815
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0659
##
```

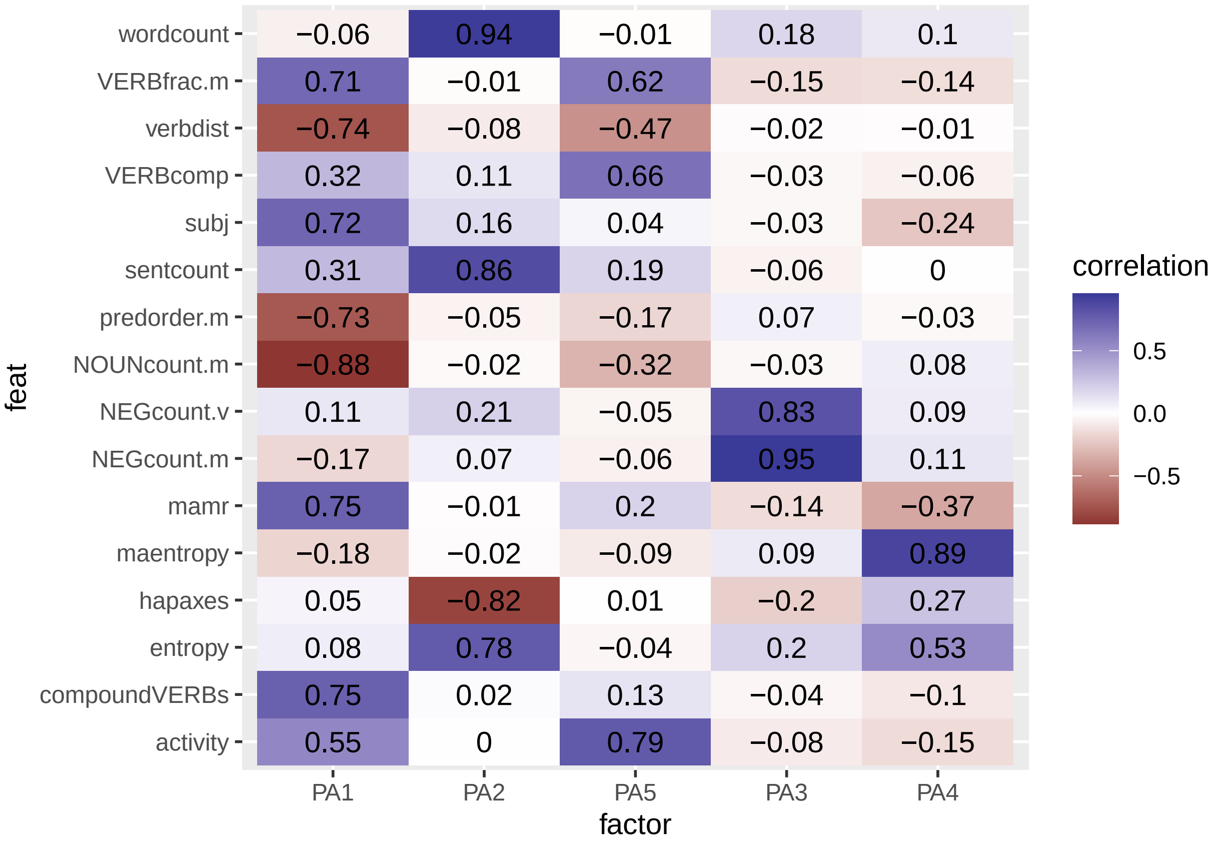
```
## Test for the significance of differences in Bindingness over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 22.2155, df = 1, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## ---------+-----------
##     TRUE |  -4.713330
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 =  0.0295
##
##   factor kruskal_p epsilon2
## 1    PA1  5.60e-78 0.465000
## 2    PA2  4.59e-01 0.000729
## 3    PA5  5.11e-23 0.130000
## 4    PA3  1.91e-12 0.065900
## 5    PA4  2.44e-06 0.029500
##
## 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

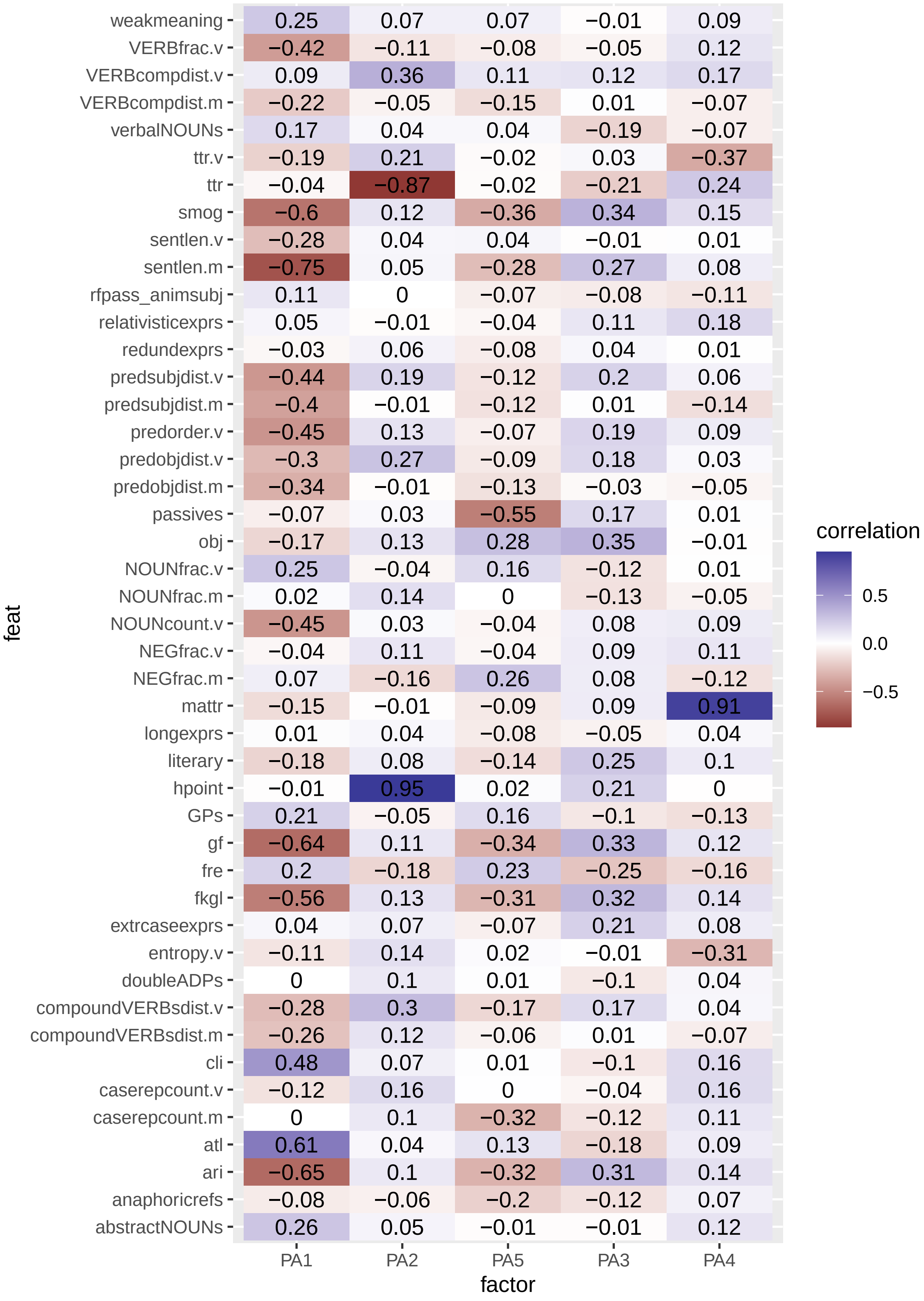
```r
data_factors_longer <- data_factors_long %>%
  pivot_longer(
    abstractNOUNs:GPs,
    names_to = "feat", values_to = "feat_value"
  )

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.
```

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



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



feat
factor
PA1
PA2
PA5
PA3
PA4
weakmeaning 0.25 0.07 0.07 −0.01 0.09
VERBfrac.v −0.42 −0.11 −0.08 −0.05 0.12
VERBcompdist.v 0.09 0.36 0.11 0.12 0.17
VERBcompdist.m −0.22 −0.05 −0.15 0.01 −0.07
verbalNOUNs 0.17 0.04 0.04 −0.19 −0.07
ttr.v −0.19 0.21 −0.02 0.03 −0.37
ttr −0.04 −0.87 −0.02 −0.21 0.24
smog −0.6 0.12 −0.36 0.34 0.15
sentlen.v −0.28 0.04 0.04 −0.01 0.01
sentlen.m −0.75 0.05 −0.28 0.27 0.08
rfpass_animsubj 0.11 0 −0.07 −0.08 −0.11
relativisticexprs 0.05 −0.01 −0.04 0.11 0.18
redundexprs −0.03 0.06 −0.08 0.04 0.01
predsubjdist.v −0.44 0.19 −0.12 0.2 0.06
predsubjdist.m −0.4 −0.01 −0.12 0.01 −0.14
predorder.v −0.45 0.13 −0.07 0.19 0.09
predobjdist.v −0.3 0.27 −0.09 0.18 0.03
predobjdist.m −0.34 −0.01 −0.13 −0.03 −0.05
passives −0.07 0.03 −0.55 0.17 0.01
obj −0.17 0.13 0.28 0.35 −0.01
NOUNfrac.v 0.25 −0.04 0.16 −0.12 0.01
NOUNfrac.m 0.02 0.14 0 −0.13 −0.05
NOUNcount.v −0.45 0.03 −0.04 0.08 0.09
NEGfrac.v −0.04 0.11 −0.04 0.09 0.11
NEGfrac.m 0.07 −0.16 0.26 0.08 −0.12
mattr −0.15 −0.01 −0.09 0.09 0.91
longexprs 0.01 0.04 −0.08 −0.05 0.04
literary −0.18 0.08 −0.14 0.25 0.1
hpoint −0.01 0.95 0.02 0.21 0
GPs 0.21 −0.05 0.16 −0.1 −0.13
gf −0.64 0.11 −0.34 0.33 0.12
fre 0.2 −0.18 0.23 −0.25 −0.16
fkgl −0.56 0.13 −0.31 0.32 0.14
extrcaseexprs 0.04 0.07 −0.07 0.21 0.08
entropy.v −0.11 0.14 0.02 −0.01 −0.31
doubleADPs 0 0.1 0.01 −0.1 0.04
compoundVERBsdist.v −0.28 0.3 −0.17 0.17 0.04
compoundVERBsdist.m −0.26 0.12 −0.06 0.01 −0.07
cli 0.48 0.07 0.01 −0.1 0.16
caserepcount.v −0.12 0.16 0 −0.04 0.16
caserepcount.m 0 0.1 −0.32 −0.12 0.11
atl 0.61 0.04 0.13 −0.18 0.09
ari −0.65 0.1 −0.32 0.31 0.14
anaphoricrefs −0.08 −0.06 −0.2 −0.12 0.07
abstractNOUNs 0.26 0.05 −0.01 −0.01 0.12
correlation
0.5
0.0
−0.5