

EFA

```
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
library(corrplot)

## corrplot 0.95 loaded
library(QuantPsyc) # for the multivariate normality test

## Loading required package: boot
## Loading required package: dplyr
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:igraph':
##
##      as_data_frame, groups, union
## The following objects are masked from 'package:stats':
##
##      filter, lag
## The following objects are masked from 'package:base':
##
##      intersect, setdiff, setequal, union
## Loading required package: purrr
##
## Attaching package: 'purrr'
## The following objects are masked from 'package:igraph':
##
##      compose, simplify
## Loading required package: MASS
##
## Attaching package: 'MASS'
```

```

## The following object is masked from 'package:dplyr':
##
##   select
##
## Attaching package: 'QuantPsyc'
## The following object is masked from 'package:base':
##
##   norm
library(dunn.test)
library(nFactors) # for the scree plot

## Loading required package: lattice
##
## Attaching package: 'lattice'
## The following object is masked from 'package:boot':
##
##   melanoma
##
## Attaching package: 'nFactors'
## The following object is masked from 'package:lattice':
##
##   parallel
library(psych) # for PA FA

##
## Attaching package: 'psych'
## The following object is masked from 'package:boot':
##
##   logit
## The following object is masked from 'package:rcompanion':
##
##   phi
library(caret) # highly correlated features removal

## Loading required package: ggplot2
##
## Attaching package: 'ggplot2'
## The following objects are masked from 'package:psych':
##
##   %+%, alpha
##
## Attaching package: 'caret'
## The following object is masked from 'package:purrr':
##
##   lift
library(tidyverse)

```

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

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

library(paletteer) # color palettes

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

## [conflicted] Will prefer dplyr::select over any other package.
conflict_prefer("filter", "dplyr")

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

Helpers

```
analyze_distributions <- function(data_factors_long, variable) {
  print(table(data_factors_long[[variable]], useNA = "ifany"))

  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]],
  alt = 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_factor_bind <- function(data, fa_fit) {
  data_factors <- bind_cols(data, fa_fit$scores %>% as.data.frame())
  colnames(data_factors) <- prettify_feat_name_vector(colnames(data_factors))

  fnames <- colnames(fa_fit$loadings)

  data_factors_long <- data_factors %>%
    pivot_longer(
      any_of(fnames),
      names_to = "factor", values_to = "factor_score"
    ) %>%
    mutate(across(
      factor,
      ~ factor(.x, levels = fnames)
    )) %>%
    select(
      all_of(1:(.firstnonmetacolumn - 1)), factor, factor_score, everything()
    )

  data_factors_longer <- data_factors_long %>% pivot_longer(
    all_of((.firstnonmetacolumn + 2):ncol(data_factors_long)),

```

```

    names_to = "feat", values_to = "feat_value"
  )

  return(list(
    data = data_factors,
    long = data_factors_long,
    feat_long = data_factors_longer
  ))
}

```

Load and tidy data

```

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

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

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.
.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,
RuleWrongVerbominalCase = 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,
    RuleWrongVerbNomininalCase,
    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>, ...
colnames(data_clean) <- prettify_feat_name_vector(colnames(data_clean))

```

Important features identification

```

feature_importances <- read_csv("../importance_measures/featcomp.csv")

## Rows: 61 Columns: 21
## -- Column specification -----
## Delimiter: ","
## chr  (2): Variable, Sign
## dbl (15): Importance, p_value, estimate, wilcox_p, wilcox_r, kw_p, kw_chi2, ...
## lgl  (4): selected_pval, wilcox_sel, kw_sel, selected_reg
##
## 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.
selected_features_names <- feature_importances %>%
  filter(kw_sel) %>%
  pull(Variable)
```

Correlations

See Levshina (2015: 353-54).

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

  cor_tibble_long <- cor_matrix %>%
    as_tibble() %>%
    mutate(featl = rownames(cor_matrix)) %>%
    pivot_longer(!featl, 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(featl = rownames(cor_matrix)) %>%
    pivot_longer(!featl, names_to = "feat2", values_to = "cor") %>%
    mutate(abs_cor = abs(cor)) %>%
    filter(featl != 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 %>%
  # remove readability metrics as they're conceptually different
  # to the remaining features
  select(!c(ari, cli, fkg1, fre, gf, smog)) %>%
  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(feet1 != feat2 & abs_cor > .hcorrcutoff) %>%
arrange(feet1, -abs_cor) %>%
print(n = 100)
```

```
## # A tibble: 4 x 4
##   feat1    feat2      cor abs_cor
##   <chr>    <chr>    <dbl> <dbl>
## 1 hpoint  wordcount 0.958  0.958
## 2 maentropy mattr    0.964  0.964
## 3 mattr   maentropy 0.964  0.964
## 4 wordcount 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/ fkg1 almost 0.95, but fkg1 coefficients adjusted for Czech are available
- **atl**: corr. w/ cli around 0.96; unlike cli, atl is not a readability metric

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

```
## Compare row 7 and column 6 with corr 0.958
## Means: 0.179 vs 0.186 so flagging column 6
## Compare row 20 and column 15 with corr 0.964
## Means: 0.166 vs 0.187 so flagging column 15
## All correlations <= 0.9
```

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

```
## [1] "hpoint" "mattr"
```

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

```
analyze_correlation(data_pureish_striphigh)$cor_tibble_long %>%
  filter(feet1 != feat2 & abs_cor > .hcorrcutoff) %>%
  arrange(feet1, -abs_cor) %>%
  print(n = 100)
```

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

Low correlations

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

```
low_correlating_features <- analyze_correlation(data_pureish_striphigh)$
  cor_tibble_long %>%
  filter(feet1 != feat2) %>%
  group_by(feet1) %>%
```

```

summarize(max_cor = max(abs_cor)) %>%
filter(max_cor < .lcorrcutoff) %>%
pull(featl)

feature_importances %>%
  filter(Variable %in% low_correlating_features) %>%
  pull(Variable)

## [1] "anaphoricrefs"      "extrcaseexprs"      "caserepcount.v"
## [4] "redundexprs"        "relativisticexprs"  "VERBcompdist.m"
## [7] "NOUNfrac.v"         "abstractNOUNs"

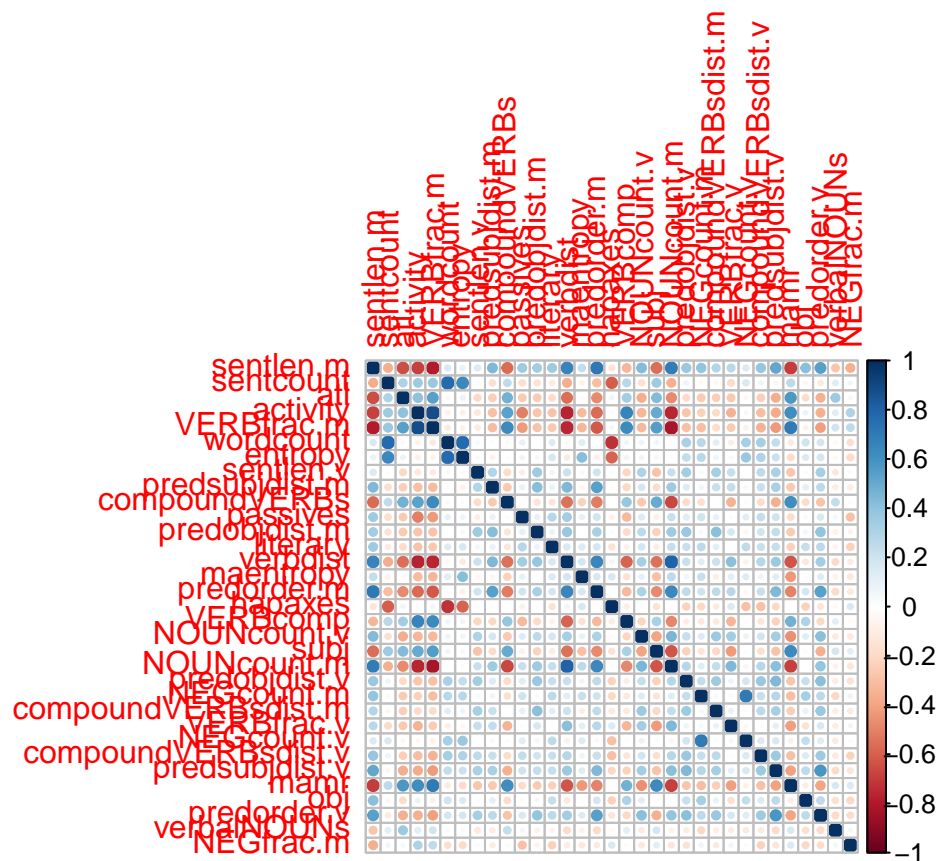
data_pure <- data_pureish_striphigh %>%
  select(!any_of(low_correlating_features))

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

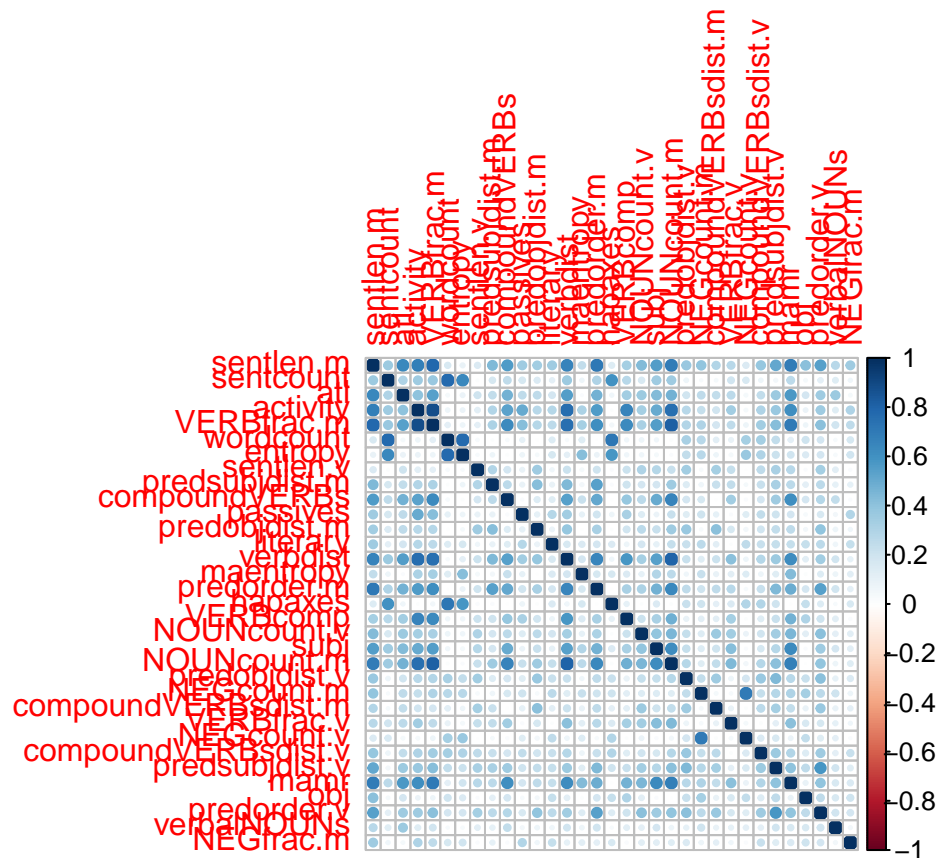
```

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)

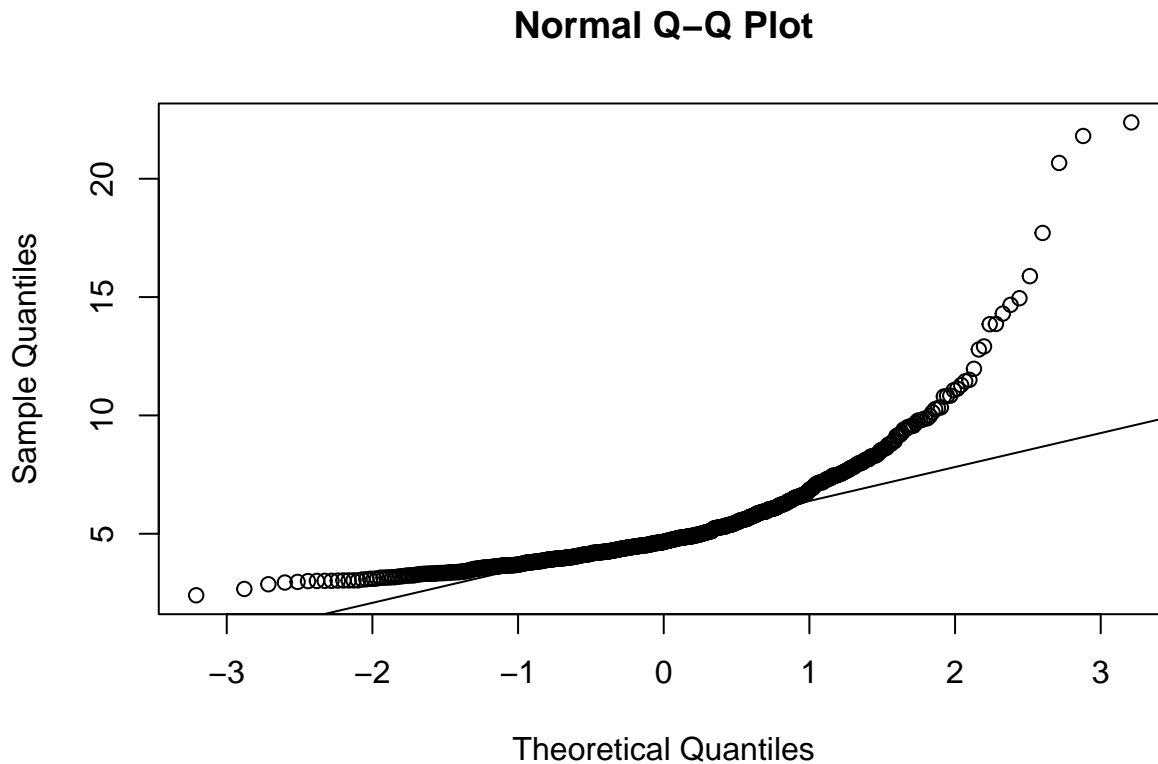
network <- graph_from_data_frame(
  network_edges,
  directed = FALSE
)
E(network)$weight <- network_edges$abs_cor
network_communities <- cluster_optimal(network)

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
)
```



```
mardia(data_scaled)
```



```
## Call: mardia(x = data_scaled)
##
## Mardia tests of multivariate skew and kurtosis
## Use describe(x) the to get univariate tests
## n.obs = 753  num.vars = 33
## b1p = 1072.73  skew = 134627.8  with probability <= 0
## small sample skew = 135195.8  with probability <= 0
## b2p = 2721.15  kurtosis = 447.09  with probability <= 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.

Bad, yet broad FA

No. of vectors

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

```
## Parallel analysis suggests that the number of factors = 8 and the number of components = NA
dev.off()
```

```
## pdf
## 2
```

Model

```
set.seed(42)
```

```
fa_broad <- fa(  
  data_scaled,  
  nfactors = 8,  
  fm = "pa",  
  rotate = "promax",  
  oblique.scores = TRUE,  
  scores = "tenBerge",  
  n.iter = 100  
)
```

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

```
fa_broad
```

```
## 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  PA3  PA5  PA6  PA4  PA8  PA7  h2  u2  
## sentlen.m    -0.62 -0.02 -0.03 -0.28  0.00  0.37  0.15 -0.02 0.94 0.063  
## sentcount     0.15  0.96  0.03  0.32 -0.07 -0.16  0.00 -0.01 0.93 0.066  
## atl           0.70  0.00 -0.02  0.06 -0.05 -0.13  0.10  0.30 0.57 0.431  
## activity      0.66 -0.01  0.10  0.47  0.00  0.31 -0.09 -0.09 0.89 0.106  
## VERBfrac.m    0.80 -0.06  0.20  0.35 -0.02  0.10 -0.12 -0.05 0.90 0.100  
## wordcount    -0.15  0.95  0.00  0.01  0.02  0.00 -0.05  0.01 0.89 0.114  
## entropy       0.03  0.72  0.07 -0.02  0.10 -0.04 -0.12  0.39 0.86 0.141  
## sentlen.v     0.00 -0.01  0.73  0.28  0.01 -0.15  0.05 -0.02 0.46 0.538  
## predsubjdist.m -0.08 -0.04  0.25  0.12 -0.04  0.06  0.55 -0.04 0.45 0.555  
## compoundVERBs  0.99 -0.15  0.30 -0.31  0.07 -0.18 -0.14 -0.04 0.70 0.298  
## passives      0.03 -0.09 -0.03 -0.79  0.15 -0.25 -0.06 -0.09 0.57 0.427  
## predobjdist.m  0.08 -0.12  0.60  0.01 -0.05 -0.08  0.29  0.00 0.42 0.583  
## literary      0.00 -0.04  0.07 -0.34  0.15  0.14 -0.05  0.06 0.24 0.758  
## verbdist     -0.74  0.00  0.00 -0.12 -0.06 -0.25  0.26 -0.04 0.81 0.188  
## maentropy    -0.19 -0.07 -0.15 -0.03  0.12 -0.01 -0.01  0.82 0.76 0.245  
## predorder.m  -0.45 -0.07  0.06  0.06 -0.04  0.19  0.51  0.07 0.70 0.297  
## hapaxes       0.10 -0.83  0.07  0.07  0.01 -0.10  0.01  0.29 0.72 0.282  
## VERBcomp      0.56  0.02 -0.01  0.15 -0.15  0.54 -0.01  0.04 0.60 0.404  
## NOUNcount.v   -0.33 -0.04  0.43 -0.08 -0.05  0.01 -0.22 -0.03 0.41 0.594  
## subj          0.69  0.12 -0.14 -0.04  0.11 -0.02  0.13 -0.14 0.58 0.422  
## NOUNcount.m   -0.84  0.05  0.01 -0.08 -0.17 -0.10  0.14  0.07 0.79 0.209  
## predobjdist.v  0.05  0.14  0.51 -0.07  0.07  0.04  0.07  0.02 0.39 0.606  
## NEGcount.m    0.04 -0.05 -0.06  0.08  1.00  0.08  0.03  0.09 0.94 0.063  
## compoundVERBsdist.m 0.13 -0.02  0.71 -0.14 -0.08 -0.04 -0.03 -0.14 0.43 0.566  
## VERBfrac.v    -0.55 -0.03  0.15  0.23 -0.04 -0.21 -0.06  0.06 0.35 0.648  
## NEGcount.v    0.21  0.09  0.01 -0.03  0.75  0.02 -0.11  0.07 0.59 0.415  
## compoundVERBsdist.v -0.07  0.23  0.28 -0.20  0.04  0.00  0.06 -0.03 0.33 0.672  
## predsubjdist.v -0.14  0.10  0.38 -0.03  0.10  0.13  0.17  0.03 0.47 0.533  
## mamr          0.84 -0.07 -0.06  0.02  0.01  0.02  0.16 -0.17 0.77 0.234  
## obj           0.08 -0.03 -0.06  0.00  0.08  0.83  0.10 -0.02 0.68 0.322
```



```

## predorder.v      -0.05 -0.02  0.52 -0.05  0.07  0.16  0.17  0.08 0.54 0.463
## verbalNOUNs      0.23  0.05 -0.02 -0.12 -0.14 -0.18  0.00  0.04 0.14 0.862
## NEGfrac.m        -0.03 -0.02 -0.03  0.60  0.29 -0.21  0.09 -0.09 0.40 0.602
##                  com
## sentlen.m         2.2
## sentcount         1.3
## atl               1.5
## activity          2.4
## VERBfrac.m        1.6
## wordcount         1.1
## entropy           1.7
## sentlen.v         1.4
## predsubjdist.m    1.6
## compoundVERBs      1.6
## passives          1.4
## predobjdist.m     1.6
## literary          2.0
## verbdist          1.6
## maentropy         1.2
## predorder.m       2.4
## hapaxes           1.3
## VERBcomp          2.3
## NOUNcount.v       2.6
## subj             1.4
## NOUNcount.m       1.2
## predobjdist.v     1.3
## NEGcount.m        1.1
## compoundVERBsdist.m 1.3
## VERBfrac.v        1.9
## NEGcount.v        1.3
## compoundVERBsdist.v 3.1
## predsubjdist.v    2.4
## mamr              1.2
## obj               1.1
## predorder.v       1.6
## verbalNOUNs       3.4
## NEGfrac.m         1.9
##
##                  PA1  PA2  PA3  PA5  PA6  PA4  PA8  PA7
## SS loadings       6.71 3.10 2.53 2.08 1.74 1.56 1.29 1.19
## Proportion Var    0.20 0.09 0.08 0.06 0.05 0.05 0.04 0.04
## Cumulative Var    0.20 0.30 0.37 0.44 0.49 0.54 0.58 0.61
## Proportion Explained 0.33 0.15 0.13 0.10 0.09 0.08 0.06 0.06
## Cumulative Proportion 0.33 0.49 0.61 0.71 0.80 0.88 0.94 1.00
##
## With factor correlations of
##      PA1  PA2  PA3  PA5  PA6  PA4  PA8  PA7
## PA1  1.00  0.11 -0.56  0.38 -0.37 -0.18 -0.36 -0.17
## PA2  0.11  1.00  0.17 -0.26  0.27  0.25  0.01  0.18
## PA3 -0.56  0.17  1.00 -0.33  0.30  0.32  0.24  0.11
## PA5  0.38 -0.26 -0.33  1.00 -0.34 -0.23 -0.38 -0.17
## PA6 -0.37  0.27  0.30 -0.34  1.00  0.32  0.11  0.07
## PA4 -0.18  0.25  0.32 -0.23  0.32  1.00  0.00  0.08
## PA8 -0.36  0.01  0.24 -0.38  0.11  0.00  1.00 -0.10

```

```

## PA7 -0.17  0.18  0.11 -0.17  0.07  0.08 -0.10  1.00
##
## Mean item complexity = 1.7
## Test of the hypothesis that 8 factors are sufficient.
##
## df null model = 528 with the objective function = 24.21 with Chi Square = 17922.49
## df of the model are 292 and the objective function was 2.94
##
## 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 514.88 with prob < 1.6e-14
## The total n.obs was 753 with Likelihood Chi Square = 2157.52 with prob < 2.7e-281
##
## Tucker Lewis Index of factoring reliability = 0.805
## RMSEA index = 0.092 and the 90 % confidence intervals are 0.089 0.096
## BIC = 223.3
## Fit based upon off diagonal values = 0.99
## Measures of factor score adequacy
##
## Correlation of (regression) scores with factors PA1 PA2 PA3 PA5 PA6 PA4
## Multiple R square of scores with factors 0.98 0.98 0.92 0.94 0.98 0.94
## Minimum correlation of possible factor scores 0.96 0.96 0.85 0.89 0.96 0.89
## Correlation of (regression) scores with factors PA8 PA7
## Multiple R square of scores with factors 0.92 0.92 0.70 0.77 0.91 0.78
## Minimum correlation of possible factor scores 0.87 0.91
## Multiple R square of scores with factors 0.75 0.82
## Minimum correlation of possible factor scores 0.50 0.65
##
## Coefficients and bootstrapped confidence intervals
##
## low PA1 upper low PA2 upper low PA3 upper low
## sentlen.m -0.71 -0.62 -0.52 -0.06 -0.02 0.01 -0.09 -0.03 0.04 -0.33
## sentcount 0.09 0.15 0.20 0.91 0.96 1.01 -0.01 0.03 0.07 0.24
## atl 0.52 0.70 0.76 -0.06 0.00 0.08 -0.10 -0.02 0.07 -0.07
## activity 0.55 0.66 0.77 -0.05 -0.01 0.03 0.04 0.10 0.15 0.40
## VERBfrac.m 0.66 0.80 0.94 -0.09 -0.06 -0.01 0.13 0.20 0.25 0.25
## wordcount -0.18 -0.15 -0.09 0.91 0.95 0.98 -0.04 0.00 0.04 -0.03
## entropy -0.04 0.03 0.07 0.69 0.72 0.76 0.01 0.07 0.11 -0.08
## sentlen.v -0.11 0.00 0.08 -0.06 -0.01 0.06 0.57 0.73 0.91 0.19
## predsubjdist.m -0.28 -0.08 0.03 -0.09 -0.04 0.03 0.17 0.25 0.36 -0.05
## compoundVERBs 0.81 0.99 1.15 -0.21 -0.15 -0.08 0.18 0.30 0.39 -0.38
## passives -0.05 0.03 0.11 -0.14 -0.09 -0.04 -0.11 -0.03 0.03 -0.84
## predobjdist.m -0.05 0.08 0.18 -0.18 -0.12 -0.05 0.45 0.60 0.79 -0.14
## literary -0.10 0.00 0.10 -0.12 -0.04 0.03 -0.03 0.07 0.17 -0.41
## verbdist -0.87 -0.74 -0.63 -0.04 0.00 0.03 -0.04 0.00 0.06 -0.27
## maentropy -0.32 -0.19 -0.14 -0.10 -0.07 -0.01 -0.23 -0.15 -0.09 -0.13
## predorder.m -0.74 -0.45 -0.30 -0.10 -0.07 -0.01 -0.02 0.06 0.17 -0.15
## hapaxes -0.01 0.10 0.16 -0.86 -0.83 -0.77 0.00 0.07 0.12 -0.01
## VERBcomp 0.45 0.56 0.65 -0.03 0.02 0.07 -0.08 -0.01 0.06 0.07
## NOUNcount.v -0.42 -0.33 -0.16 -0.13 -0.04 0.03 0.27 0.43 0.58 -0.13
## subj 0.55 0.69 0.77 0.07 0.12 0.18 -0.21 -0.14 -0.07 -0.15
## NOUNcount.m -1.00 -0.84 -0.69 0.00 0.05 0.10 -0.05 0.01 0.09 -0.14
## predobjdist.v -0.12 0.05 0.20 0.04 0.14 0.24 0.38 0.51 0.67 -0.19
## NEGcount.m -0.04 0.04 0.08 -0.09 -0.05 0.00 -0.12 -0.06 -0.01 -0.02
## compoundVERBsdist.m 0.00 0.13 0.25 -0.09 -0.02 0.06 0.56 0.71 0.85 -0.21

```

## VERBfrac.v	-0.66	-0.55	-0.41	-0.11	-0.03	0.06	0.06	0.15	0.26	0.14
## NEGcount.v	0.15	0.21	0.30	0.03	0.09	0.13	-0.04	0.01	0.06	-0.09
## compoundVERBsdist.v	-0.19	-0.07	0.05	0.15	0.23	0.31	0.15	0.28	0.43	-0.30
## predsubjdist.v	-0.30	-0.14	-0.03	0.04	0.10	0.18	0.25	0.38	0.52	-0.15
## mamr	0.69	0.84	0.91	-0.12	-0.07	-0.01	-0.12	-0.06	0.02	-0.08
## obj	0.01	0.08	0.14	-0.06	-0.03	0.02	-0.12	-0.06	0.00	-0.06
## predorder.v	-0.20	-0.05	0.06	-0.09	-0.02	0.06	0.36	0.52	0.70	-0.16
## verbalNOUNs	0.13	0.23	0.31	-0.02	0.05	0.12	-0.12	-0.02	0.06	-0.20
## NEGfrac.m	-0.14	-0.03	0.04	-0.08	-0.02	0.04	-0.12	-0.03	0.06	0.47
##	PA5	upper	low	PA6	upper	low	PA4	upper	low	PA8
## sentlen.m	-0.28	-0.22	-0.04	0.00	0.05	0.31	0.37	0.44	0.05	0.15
## sentcount	0.32	0.35	-0.11	-0.07	-0.04	-0.21	-0.16	-0.12	-0.18	0.00
## atl	0.06	0.12	-0.14	-0.05	0.03	-0.25	-0.13	-0.06	-0.39	0.10
## activity	0.47	0.54	-0.05	0.00	0.04	0.26	0.31	0.39	-0.22	-0.09
## VERBfrac.m	0.35	0.43	-0.07	-0.02	0.03	0.05	0.10	0.17	-0.41	-0.12
## wordcount	0.01	0.06	-0.01	0.02	0.06	-0.04	0.00	0.04	-0.13	-0.05
## entropy	-0.02	0.02	0.06	0.10	0.15	-0.08	-0.04	0.00	-0.39	-0.12
## sentlen.v	0.28	0.35	-0.07	0.01	0.07	-0.22	-0.15	-0.09	-0.08	0.05
## predsubjdist.m	0.12	0.22	-0.14	-0.04	0.05	-0.07	0.06	0.20	0.13	0.55
## compoundVERBs	-0.31	-0.20	0.01	0.07	0.15	-0.26	-0.18	-0.10	-0.47	-0.14
## passives	-0.79	-0.66	0.08	0.15	0.22	-0.32	-0.25	-0.18	-0.15	-0.06
## predobjdist.m	0.01	0.12	-0.17	-0.05	0.05	-0.17	-0.08	0.02	-0.06	0.29
## literary	-0.34	-0.22	0.07	0.15	0.26	0.06	0.14	0.23	-0.16	-0.05
## verbdist	-0.12	-0.02	-0.11	-0.06	-0.02	-0.31	-0.25	-0.21	0.17	0.26
## maentropy	-0.03	0.02	0.07	0.12	0.17	-0.07	-0.01	0.03	-0.43	-0.01
## predorder.m	0.06	0.15	-0.17	-0.04	0.07	0.06	0.19	0.30	0.22	0.51
## hapaxes	0.07	0.12	-0.06	0.01	0.06	-0.16	-0.10	-0.05	-0.19	0.01
## VERBcomp	0.15	0.22	-0.22	-0.15	-0.07	0.46	0.54	0.65	-0.23	-0.01
## NOUNcount.v	-0.08	0.07	-0.12	-0.05	0.06	-0.06	0.01	0.10	-0.40	-0.22
## subj	-0.04	0.03	0.02	0.11	0.17	-0.08	-0.02	0.05	-0.08	0.13
## NOUNcount.m	-0.08	0.00	-0.25	-0.17	-0.10	-0.18	-0.10	-0.03	0.03	0.14
## predobjdist.v	-0.07	0.04	-0.03	0.07	0.18	-0.08	0.04	0.15	-0.20	0.07
## NEGcount.m	0.08	0.13	0.83	1.00	1.16	0.04	0.08	0.15	-0.13	0.03
## compoundVERBsdist.m	-0.14	-0.03	-0.15	-0.08	-0.01	-0.11	-0.04	0.03	-0.16	-0.03
## VERBfrac.v	0.23	0.35	-0.12	-0.04	0.05	-0.31	-0.21	-0.12	-0.39	-0.06
## NEGcount.v	-0.03	0.04	0.64	0.75	0.93	-0.04	0.02	0.09	-0.31	-0.11
## compoundVERBsdist.v	-0.20	-0.09	-0.04	0.04	0.13	-0.10	0.00	0.11	-0.08	0.06
## predsubjdist.v	-0.03	0.07	-0.02	0.10	0.20	0.05	0.13	0.23	-0.10	0.17
## mamr	0.02	0.06	-0.09	0.01	0.06	-0.05	0.02	0.09	-0.09	0.16
## obj	0.00	0.06	0.02	0.08	0.18	0.74	0.83	0.97	-0.02	0.10
## predorder.v	-0.05	0.03	-0.03	0.07	0.16	0.08	0.16	0.27	-0.04	0.17
## verbalNOUNs	-0.12	-0.02	-0.24	-0.14	-0.03	-0.31	-0.18	-0.08	-0.21	0.00
## NEGfrac.m	0.60	0.67	0.18	0.29	0.40	-0.30	-0.21	-0.14	-0.17	0.09
##	upper	low	PA7	upper						
## sentlen.m	0.45	-0.06	-0.02	0.05						
## sentcount	0.05	-0.07	-0.01	0.02						
## atl	0.31	0.13	0.30	0.45						
## activity	-0.02	-0.16	-0.09	-0.04						
## VERBfrac.m	0.03	-0.12	-0.05	0.01						
## wordcount	0.01	-0.03	0.01	0.07						
## entropy	-0.03	0.33	0.39	0.57						
## sentlen.v	0.28	-0.11	-0.02	0.07						
## predsubjdist.m	1.32	-0.30	-0.04	0.15						
## compoundVERBs	0.04	-0.12	-0.04	0.05						

```

## passives          0.08 -0.16 -0.09 -0.01
## predobjdist.m     0.86 -0.16  0.00  0.09
## literary          0.11 -0.02  0.06  0.18
## verbdist          0.51 -0.12 -0.04 -0.01
## maentropy         0.11  0.68  0.82  1.07
## predorder.m       0.93 -0.12  0.07  0.12
## hapaxes           0.07  0.22  0.29  0.38
## VERBcomp          0.11 -0.04  0.04  0.12
## NOUNcount.v       0.05 -0.13 -0.03  0.17
## subj              0.25 -0.32 -0.14 -0.09
## NOUNcount.m       0.38 -0.03  0.07  0.16
## predobjdist.v     0.41 -0.11  0.02  0.12
## NEGcount.m        0.12  0.04  0.09  0.15
## compoundVERBsdist.m 0.23 -0.25 -0.14 -0.05
## VERBfrac.v        0.31 -0.04  0.06  0.24
## NEGcount.v        0.00  0.02  0.07  0.19
## compoundVERBsdist.v 0.30 -0.14 -0.03  0.06
## predsubjdist.v    0.56 -0.11  0.03  0.12
## mamr              0.26 -0.37 -0.17 -0.11
## obj               0.31 -0.09 -0.02  0.03
## predorder.v       0.45 -0.04  0.08  0.18
## verbalNOUNs       0.15 -0.11  0.04  0.18
## NEGfrac.m         0.30 -0.24 -0.09 -0.03
##
## Interfactor correlations and bootstrapped confidence intervals
##      lower estimate upper
## PA1-PA2 -0.326   0.1108 0.449
## PA1-PA3 -1.004  -0.5622 0.120
## PA1-PA5 -0.861   0.3830 0.349
## PA1-PA6 -0.797  -0.3665 0.291
## PA1-PA4 -0.578  -0.1818 0.099
## PA1-PA8 -0.565  -0.3611 0.170
## PA1-PA7 -0.449  -0.1660 0.163
## PA2-PA3 -0.041   0.1702 0.327
## PA2-PA5 -0.271  -0.2586 0.530
## PA2-PA6 -0.242   0.2683 0.517
## PA2-PA4 -0.124   0.2463 0.491
## PA2-PA8 -0.190   0.0064 0.462
## PA2-PA7 -0.149   0.1785 0.317
## PA3-PA5 -0.382  -0.3255 0.750
## PA3-PA6 -0.407   0.3000 0.768
## PA3-PA4 -0.116   0.3241 0.596
## PA3-PA8 -0.211   0.2427 0.565
## PA3-PA7 -0.236   0.1085 0.449
## PA5-PA6 -0.521  -0.3378 0.701
## PA5-PA4 -0.378  -0.2304 0.607
## PA5-PA8 -0.314  -0.3838 0.490
## PA5-PA7 -0.299  -0.1659 0.382
## PA6-PA4 -0.285   0.3221 0.557
## PA6-PA8 -0.280   0.1114 0.468
## PA6-PA7 -0.297   0.0710 0.360
## PA4-PA8 -0.236  -0.0029 0.425
## PA4-PA7 -0.236   0.0752 0.277
## PA8-PA7 -0.283  -0.1047 0.310

```

Healthiness diagnostics

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

```
## # A tibble: 33 x 2
##   feat                maxload
##   <chr>              <dbl>
## 1 verbalNOUNs        0.232
## 2 compoundVERBsdist.v 0.281
## 3 literary           0.343
## 4 predsubjdist.v     0.377
## 5 NOUNcount.v        0.431
## 6 predobjdist.v      0.509
## 7 predorder.m        0.515
## 8 predorder.v        0.519
## 9 VERBfrac.v         0.549
## 10 predsubjdist.m    0.551
## # i 23 more rows
```

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

```
##      verbalNOUNs      literary compoundVERBsdist.v      VERBfrac.v
##      0.1379713      0.2423431      0.3280540      0.3524992
##      predobjdist.v      NEGfrac.m      NOUNcount.v      predobjdist.m
##      0.3939045      0.3975483      0.4064061      0.4169727
## compoundVERBsdist.m      predsubjdist.m      sentlen.v      predsubjdist.v
##      0.4336188      0.4453023      0.4615505      0.4669617
##      predorder.v      atl      passives      subj
##      0.5370148      0.5694476      0.5733804      0.5775257
##      NEGcount.v      VERBcomp      obj      compoundVERBs
##      0.5854885      0.5958715      0.6784960      0.7020210
##      predorder.m      hapaxes      maentropy      mamr
##      0.7030408      0.7184036      0.7553256      0.7664031
##      NOUNcount.m      verbdist      entropy      wordcount
##      0.7910351      0.8118113      0.8591141      0.8864995
##      activity      VERBfrac.m      sentcount      sentlen.m
##      0.8937370      0.8998234      0.9344065      0.9365817
##      NEGcount.m
##      0.9365996
```

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

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

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

```
##          wordcount      NEGcount.m          obj          mamr
##          1.058480          1.059835          1.079227          1.183128
##          NOUNcount.m      maentropy      NEGcount.v compoundVERBsdist.m
##          1.203656          1.249629          1.261795          1.268893
##          predobjdist.v      hapaxes      sentcount          passives
##          1.333335          1.333578          1.346796          1.350058
##          subj      sentlen.v          atl      predorder.v
##          1.372625          1.381042          1.509559          1.551827
##          verbdist      compoundVERBs      VERBfrac.m      predobjdist.m
##          1.558892          1.579530          1.616498          1.633887
##          predsubjdist.m      entropy      NEGfrac.m      VERBfrac.v
##          1.647062          1.696694          1.871425          1.926064
##          literary      sentlen.m      VERBcomp      predsubjdist.v
##          1.976897          2.244205          2.308159          2.404788
##          predorder.m      activity      NOUNcount.v compoundVERBsdist.v
##          2.412118          2.434222          2.574050          3.113858
##          verbalNOUNs
##          3.371824
```

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

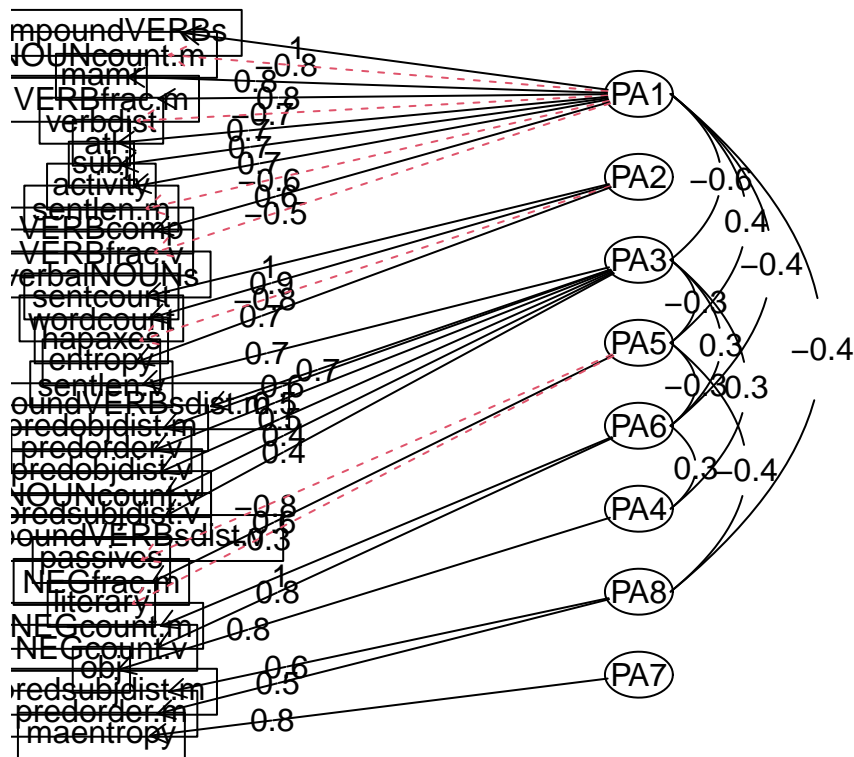
```
## [1] "sentlen.m"          "activity"          "predorder.m"
## [4] "VERBcomp"          "NOUNcount.v"      "compoundVERBsdist.v"
## [7] "predsubjdist.v"    "verbalNOUNs"
```

Loadings

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

```
fa.diagram(fa_broad)
```

Factor Analysis



fa_broad\$loadings

```
##
## Loadings:
##
```

	PA1	PA2	PA3	PA5	PA6	PA4	PA8	PA7
sentlen.m	-0.619			-0.283		0.366	0.147	
sentcount	0.152	0.961		0.317		-0.161		
at1	0.695					-0.127	0.103	0.297
activity	0.661			0.473		0.306		
VERBfrac.m	0.798		0.196	0.346		0.100	-0.120	
wordcount	-0.150	0.946						
entropy		0.717			0.102		-0.120	0.390
sentlen.v			0.731	0.275		-0.147		
predsubdist.m			0.254	0.122			0.551	
compoundVERBs	0.992	-0.154	0.296	-0.308		-0.177	-0.142	
passives				-0.790	0.146	-0.248		
predobjdist.m		-0.116	0.598				0.289	
literary				-0.343	0.149	0.136		
verbdist	-0.741			-0.118		-0.246	0.258	
maentropy	-0.190		-0.154		0.125			0.819
predorder.m	-0.452					0.188	0.515	
hapaxes	0.103	-0.829						0.286
VERBcomp	0.555			0.145	-0.151	0.538		
NOUNcount.v	-0.326		0.431				-0.222	
subj	0.693	0.118	-0.143		0.105		0.131	-0.140
NOUNcount.m	-0.839				-0.168		0.139	
predobjdist.v		0.144	0.509					
NEGcount.m					0.997			

```
## compoundVERBsdist.m 0.128          0.714 -0.139          -0.142
## VERBfrac.v          -0.549          0.150 0.229          -0.213
## NEGcount.v          0.213          0.231 0.281 -0.196          0.751          -0.111
## compoundVERBsdist.v          0.231 0.281 -0.196
## predsubjdist.v      -0.144          0.377          0.129 0.174
## mamr                0.838          0.157 -0.171
## obj                0.828
## predorder.v         0.519          0.160 0.165
## verbalNOUNs         0.232          -0.118 -0.140 -0.176
## NEGfrac.m           0.598 0.295 -0.214
##
##          PA1  PA2  PA3  PA5  PA6  PA4  PA8  PA7
## SS loadings  6.541 3.204 2.638 2.011 1.855 1.689 1.032 1.136
## Proportion Var 0.198 0.097 0.080 0.061 0.056 0.051 0.031 0.034
## Cumulative Var 0.198 0.295 0.375 0.436 0.492 0.544 0.575 0.609
```

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

  loadings <- fa_broad$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
## compoundVERBs      0.992 0.992 *****
## NOUNcount.m       -0.839 0.839 *****
## mamr              0.838 0.838 *****
## VERBfrac.m        0.798 0.798 *****
## verbdist         -0.741 0.741 *****
## at1              0.695 0.695 ****
## subj             0.693 0.693 ****
## activity          0.661 0.661 ****
## sentlen.m        -0.619 0.619 ***
## VERBcomp          0.555 0.555 ***
```



```

## VERBfrac.v          -0.549 0.549 **
## predorder.m         -0.452 0.452 **
## NOUNcount.v         -0.326 0.326 *
## verbalNOUNs         0.232 0.232
## NEGcount.v          0.213 0.213
## maentropy           -0.190 0.190
## sentcount           0.152 0.152
## wordcount           -0.150 0.150
## predsubjdist.v      -0.144 0.144
## compoundVERBsdist.m  0.128 0.128
## hapaxes             0.103 0.103
##
##
## ----- PA2 -----
## loading abs_l strng
## sentcount           0.961 0.961 *****
## wordcount           0.946 0.946 *****
## hapaxes             -0.829 0.829 *****
## entropy             0.717 0.717 *****
## compoundVERBsdist.v  0.231 0.231
## compoundVERBs       -0.154 0.154
## predobjdist.v       0.144 0.144
## subj               0.118 0.118
## predobjdist.m      -0.116 0.116
##
##
## ----- PA3 -----
## loading abs_l strng
## sentlen.v           0.731 0.731 *****
## compoundVERBsdist.m  0.714 0.714 *****
## predobjdist.m       0.598 0.598 ***
## predorder.v         0.519 0.519 **
## predobjdist.v       0.509 0.509 **
## NOUNcount.v         0.431 0.431 *
## predsubjdist.v      0.377 0.377 *
## compoundVERBs        0.296 0.296
## compoundVERBsdist.v  0.281 0.281
## predsubjdist.m      0.254 0.254
## VERBfrac.m          0.196 0.196
## maentropy           -0.154 0.154
## VERBfrac.v          0.150 0.150
## subj               -0.143 0.143
##
##
## ----- PA5 -----
## loading abs_l strng
## passives            -0.790 0.790 *****
## NEGfrac.m           0.598 0.598 ***
## activity            0.473 0.473 **
## VERBfrac.m          0.346 0.346 *
## literary            -0.343 0.343 *
## sentcount           0.317 0.317
## compoundVERBs       -0.308 0.308
## sentlen.m          -0.283 0.283

```

```

## sentlen.v          0.275 0.275
## VERBfrac.v         0.229 0.229
## compoundVERBsdist.v -0.196 0.196
## VERBcomp           0.145 0.145
## compoundVERBsdist.m -0.139 0.139
## predsubjdist.m     0.122 0.122
## verbdist           -0.118 0.118
## verbalNOUNs        -0.118 0.118
##
##
## ----- PA6 -----
##          loading abs_l strng
## NEGcount.m      0.997 0.997 *****
## NEGcount.v      0.751 0.751 *****
## NEGfrac.m       0.295 0.295
## NOUNcount.m    -0.168 0.168
## VERBcomp       -0.151 0.151
## literary        0.149 0.149
## passives        0.146 0.146
## verbalNOUNs    -0.140 0.140
## maentropy       0.125 0.125
## subj           0.105 0.105
## entropy         0.102 0.102
##
##
## ----- PA4 -----
##          loading abs_l strng
## obj            0.828 0.828 *****
## VERBcomp       0.538 0.538 **
## sentlen.m      0.366 0.366 *
## activity        0.306 0.306
## passives       -0.248 0.248
## verbdist       -0.246 0.246
## NEGfrac.m      -0.214 0.214
## VERBfrac.v     -0.213 0.213
## predorder.m    0.188 0.188
## compoundVERBs  -0.177 0.177
## verbalNOUNs    -0.176 0.176
## sentcount      -0.161 0.161
## predorder.v    0.160 0.160
## sentlen.v      -0.147 0.147
## literary        0.136 0.136
## predsubjdist.v 0.129 0.129
## atl            -0.127 0.127
## VERBfrac.m     0.100 0.100
##
##
## ----- PA8 -----
##          loading abs_l strng
## predsubjdist.m 0.551 0.551 ***
## predorder.m    0.515 0.515 **
## predobjdist.m  0.289 0.289
## verbdist       0.258 0.258
## NOUNcount.v    -0.222 0.222

```

```
## predsubjdist.v    0.174 0.174
## predorder.v      0.165 0.165
## mamr             0.157 0.157
## sentlen.m        0.147 0.147
## compoundVERBs    -0.142 0.142
## NOUNcount.m      0.139 0.139
## subj            0.131 0.131
## VERBfrac.m       -0.120 0.120
## entropy          -0.120 0.120
## NEGcount.v       -0.111 0.111
## atl             0.103 0.103
##
##
## ----- PA7 -----
##                      loading abs_l strng
## maentropy          0.819 0.819 *****
## entropy            0.390 0.390 *
## atl               0.297 0.297
## hapaxes           0.286 0.286
## mamr              -0.171 0.171
## compoundVERBsdist.m -0.142 0.142
## subj             -0.140 0.140
```

hypotheses:

Uniquenesses

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

##	sentlen.m	sentcount	atl	activity
##	0.063	0.066	0.431	0.106
##	VERBfrac.m	wordcount	entropy	sentlen.v
##	0.100	0.114	0.141	0.538
##	predsubjdist.m	compoundVERBs	passives	predobjdist.m
##	0.555	0.298	0.427	0.583
##	literary	verbdist	maentropy	predorder.m
##	0.758	0.188	0.245	0.297
##	hapaxes	VERBcomp	NOUNcount.v	subj
##	0.282	0.404	0.594	0.422
##	NOUNcount.m	predobjdist.v	NEGcount.m	compoundVERBsdist.m
##	0.209	0.606	0.063	0.566
##	VERBfrac.v	NEGcount.v	compoundVERBsdist.v	predsubjdist.v
##	0.648	0.415	0.672	0.533
##	mamr	obj	predorder.v	verbalNOUNs
##	0.234	0.322	0.463	0.862
##	NEGfrac.m			
##	0.602			

Distributions over factors

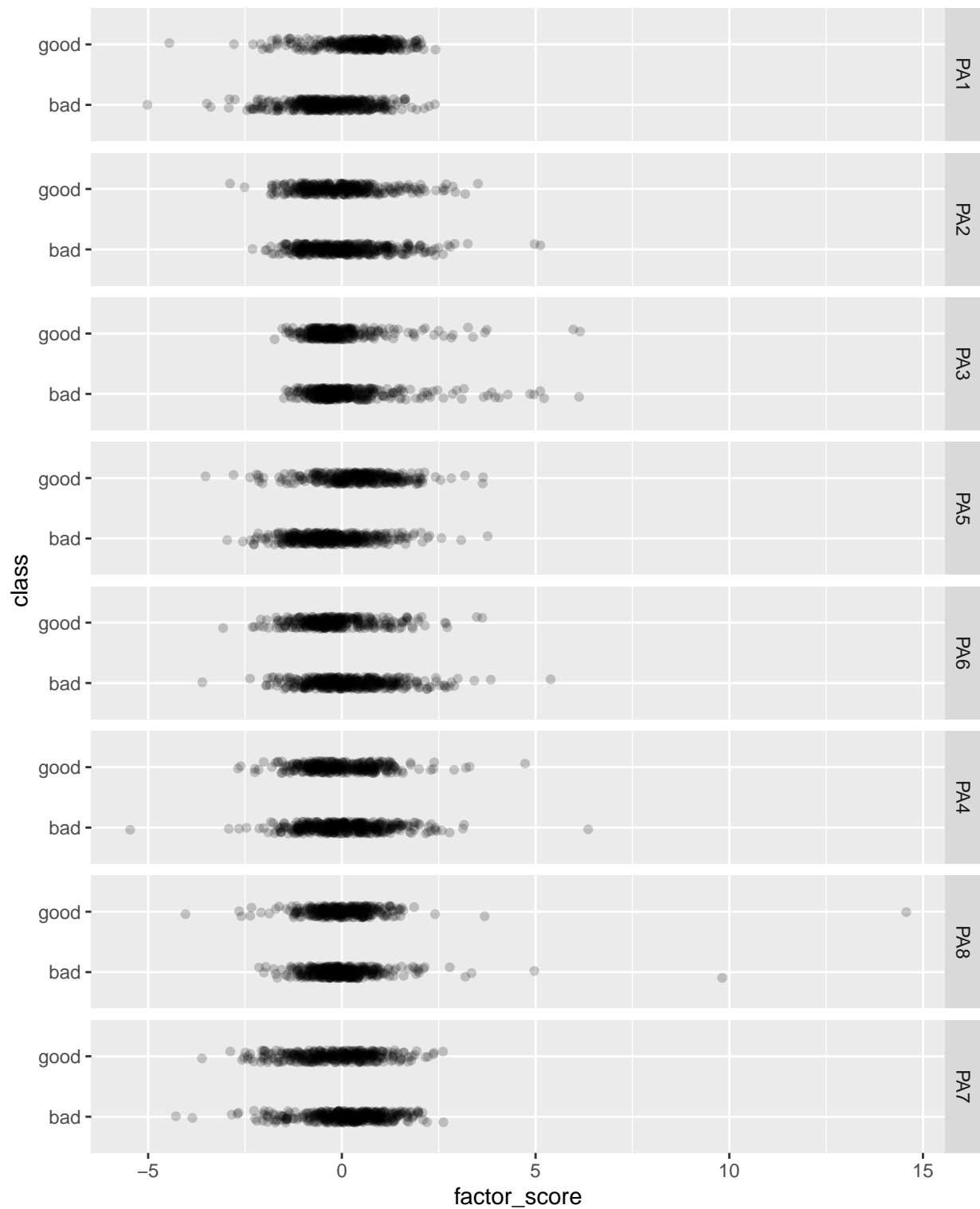
```
broad_data <- data_factor_bind(data_clean, fa_broad)
```

```
broad_data$long %>%
  group_by(factor) %>%
```

```
summarize(shapiro = shapiro.test(factor_score)$p.value)
```

```
## # A tibble: 8 x 2
##   factor shapiro
##   <fct>   <dbl>
## 1 PA1    1.41e- 8
## 2 PA2    3.52e-13
## 3 PA3    4.05e-32
## 4 PA5    1.73e- 2
## 5 PA6    7.21e-12
## 6 PA4    1.50e-12
## 7 PA8    1.34e-34
## 8 PA7    4.28e- 7
```

```
broad_data$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)
```

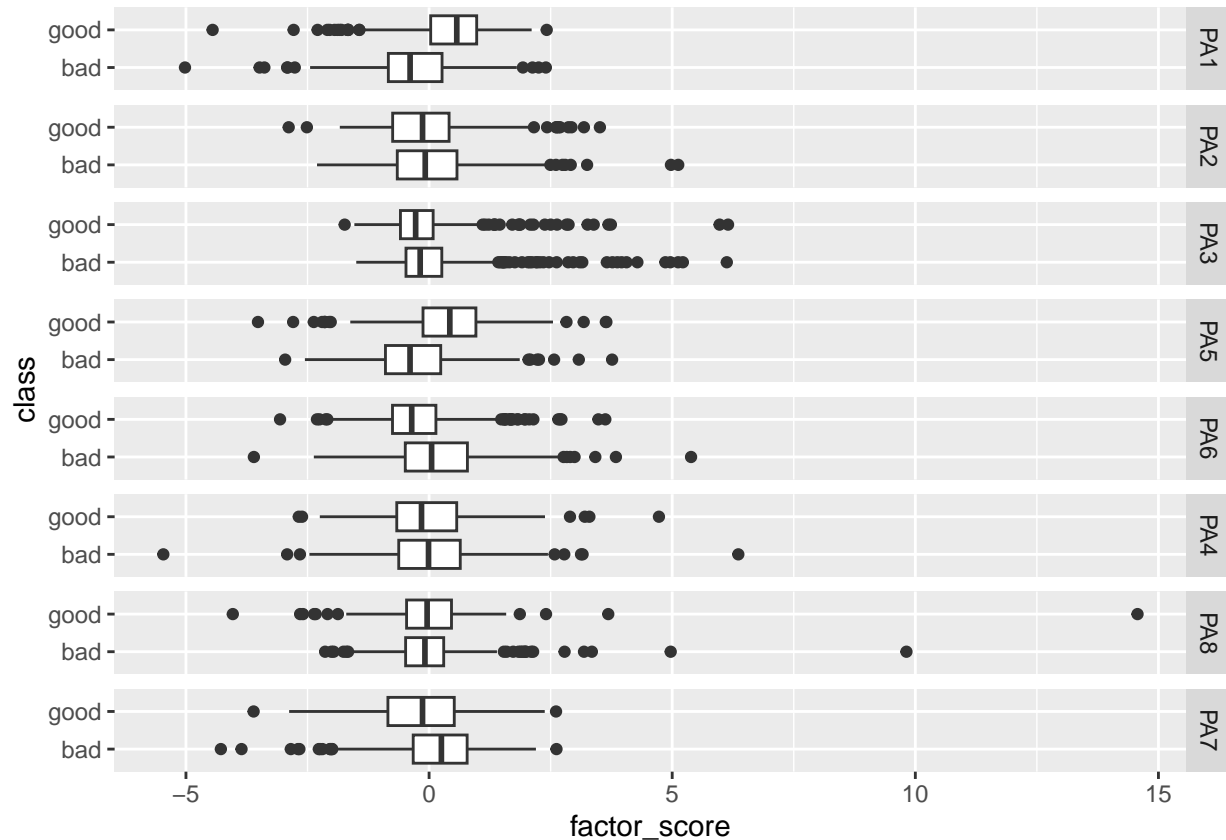


class

```
analyze_distributions(broad_data$long, "class")
```

##

```
## bad good
## 3312 2712
```



```
##
## Test for the significance of differences in class over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 134.1647, df = 1, p-value = 0
##
##
##               Comparison of x by group
##               (Bonferroni)
## Col Mean-|
## Row Mean |      bad
## -----+-----
##   good | -11.58295
##       |  0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.178
##
## Test for the significance of differences in class over PA2 :
##
##   Kruskal-Wallis rank sum test
##
```

```

## data: x and group
## Kruskal-Wallis chi-squared = 1.5495, df = 1, p-value = 0.21
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |          bad
## -----+-----
##      good |    1.244788
##          |    0.2132
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.00206
##
## Test for the significance of differences in class over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 8.5251, df = 1, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |          bad
## -----+-----
##      good |    2.919772
##          |    0.0035*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0113
##
## Test for the significance of differences in class over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 111.8462, df = 1, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |          bad
## -----+-----
##      good |   -10.57573
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha

```

```

## epsilon2 = 0.149
##
## Test for the significance of differences in class over PA6 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 35.0328, df = 1, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |          bad
## -----+-----
##    good |    5.918850
##         |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0466
##
## Test for the significance of differences in class over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 1.9676, df = 1, p-value = 0.16
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |          bad
## -----+-----
##    good |    1.402723
##         |    0.1607
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.00262
##
## Test for the significance of differences in class over PA8 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 1.7297, df = 1, p-value = 0.19
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |          bad

```



```

## -----+-----
##      good |  -1.315169
##          |    0.1885
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0023
##
## Test for the significance of differences in class over PA7 :
##
##      Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 25.6664, df = 1, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |          bad
## -----+-----
##      good |    5.066204
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0341
##
##      factor kruskal_p epsilon2
## 1      PA1 5.03e-31 0.17800
## 2      PA2 2.13e-01 0.00206
## 3      PA3 3.50e-03 0.01130
## 4      PA5 3.86e-26 0.14900
## 5      PA6 3.24e-09 0.04660
## 6      PA4 1.61e-01 0.00262
## 7      PA8 1.88e-01 0.00230
## 8      PA7 4.06e-07 0.03410
##
## p < 5e-2 found in: PA1 PA3 PA5 PA6 PA7
## p < 1e-2 found in: PA1 PA3 PA5 PA6 PA7
## p < 1e-3 found in: PA1 PA5 PA6 PA7
## p < 1e-4 found in: PA1 PA5 PA6 PA7

```

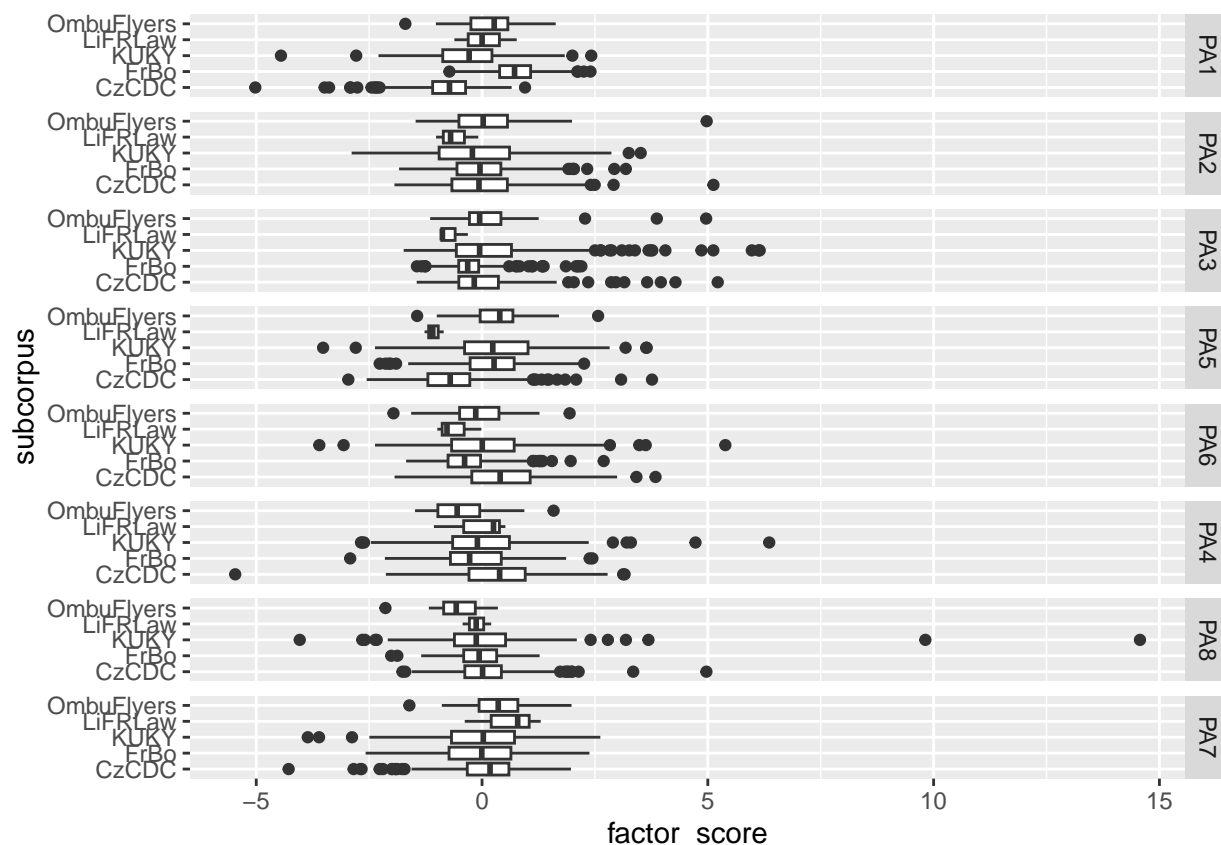
subcorpus

```
analyze_distributions(broad_data$long, "subcorpus")
```

```

##
##      CzCDC      FrBo      KUKY      LiFRLaw OmbuFlyers
##      1688      2456      1552          24          304

```



```
##
## Test for the significance of differences in subcorpus over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 395.852, df = 4, p-value = 0
##
##
##           Comparison of x by group
##           (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC      FrBo      KUKY      LiFRLaw
## -----|-----
##   FrBo | -18.96883
##         |  0.0000*
##         |
##   KUKY | -5.099316  12.96436
##         |  0.0000*  0.0000*
##         |
## LiFRLaw | -1.520822  1.399609 -0.648070
##         |  1.0000  1.0000  1.0000
##         |
## OmbuFlye | -5.887897  3.830227 -2.989708 -0.255667
##         |  0.0000*  0.0013*  0.0279*  1.0000
##
## alpha = 0.05
```

```

## Reject Ho if p <= alpha
## epsilon2 = 0.526
##
## Test for the significance of differences in subcorpus over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 5.8651, df = 4, p-value = 0.21
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC      FrBo      KUKY      LiFRLaw
## -----+-----
##   FrBo |      0.033912
##         |      1.0000
##         |
##   KUKY |      1.604396      1.706931
##         |      1.0000      0.8783
##         |
## LiFRLaw |      1.270076      1.267642      0.994997
##         |      1.0000      1.0000      1.0000
##         |
## OmbuFlye |     -0.631636     -0.664904     -1.527047     -1.416996
##         |      1.0000      1.0000      1.0000      1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0078
##
## Test for the significance of differences in subcorpus over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 32.2648, df = 4, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC      FrBo      KUKY      LiFRLaw
## -----+-----
##   FrBo |      3.399903
##         |      0.0067*
##         |
##   KUKY |     -1.106002     -4.514392
##         |      1.0000      0.0001*
##         |
## LiFRLaw |      2.014068      1.494450      2.201923
##         |      0.4400      1.0000      0.2767
##         |

```

```

## OmbuFlye | -1.595793 -3.403249 -0.965088 -2.421644
##          |      1.0000      0.0067*      1.0000      0.1545
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0429
##
## Test for the significance of differences in subcorpus over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 158.8361, df = 4, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC      FrBo      KUKY      LiFRLaw
## -----+-----
##   FrBo | -11.25575
##         |  0.0000*
##         |
##   KUKY | -9.935614  0.199036
##         |  0.0000*  1.0000
##         |
## LiFRLaw |  0.906812  2.643720  2.604923
##         |  1.0000  0.0820  0.0919
##         |
## OmbuFlye | -6.267907 -0.570014 -0.655468 -2.721021
##         |  0.0000*  1.0000  1.0000  0.0651
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.211
##
## Test for the significance of differences in subcorpus over PA6 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 93.6579, df = 4, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC      FrBo      KUKY      LiFRLaw
## -----+-----
##   FrBo |  9.518951
##         |  0.0000*
##         |
##   KUKY |  4.103058 -4.831160
##         |  0.0004*  0.0000*

```

```

##
## LiFRLaw | 2.074774 0.612126 1.372011
##          | 0.3801 1.0000 1.0000
##
## OmbuFlye | 3.100564 -1.772579 0.779434 -1.100472
##          | 0.0193* 0.7630 1.0000 1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.125
##
## Test for the significance of differences in subcorpus over PA4 :
##
## Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 52.4123, df = 4, p-value = 0
##
##
## Comparison of x by group
## (Bonferroni)
## Col Mean-|
## Row Mean | CzCDC FrBo KUKY LiFRLaw
## -----+-----
## FrBo | 6.340314
##       | 0.0000*
##
## KUKY | 3.787715 -2.073977
##       | 0.0015* 0.3808
##
## LiFRLaw | 0.760979 -0.214616 0.112935
##         | 1.0000 1.0000 1.0000
##
## OmbuFlye | 5.237906 2.070563 3.079441 0.801364
##          | 0.0000* 0.3840 0.0207* 1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0697
##
## Test for the significance of differences in subcorpus over PA8 :
##
## Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 26.9652, df = 4, p-value = 0
##
##
## Comparison of x by group
## (Bonferroni)
## Col Mean-|
## Row Mean | CzCDC FrBo KUKY LiFRLaw
## -----+-----
## FrBo | 1.444019

```

```

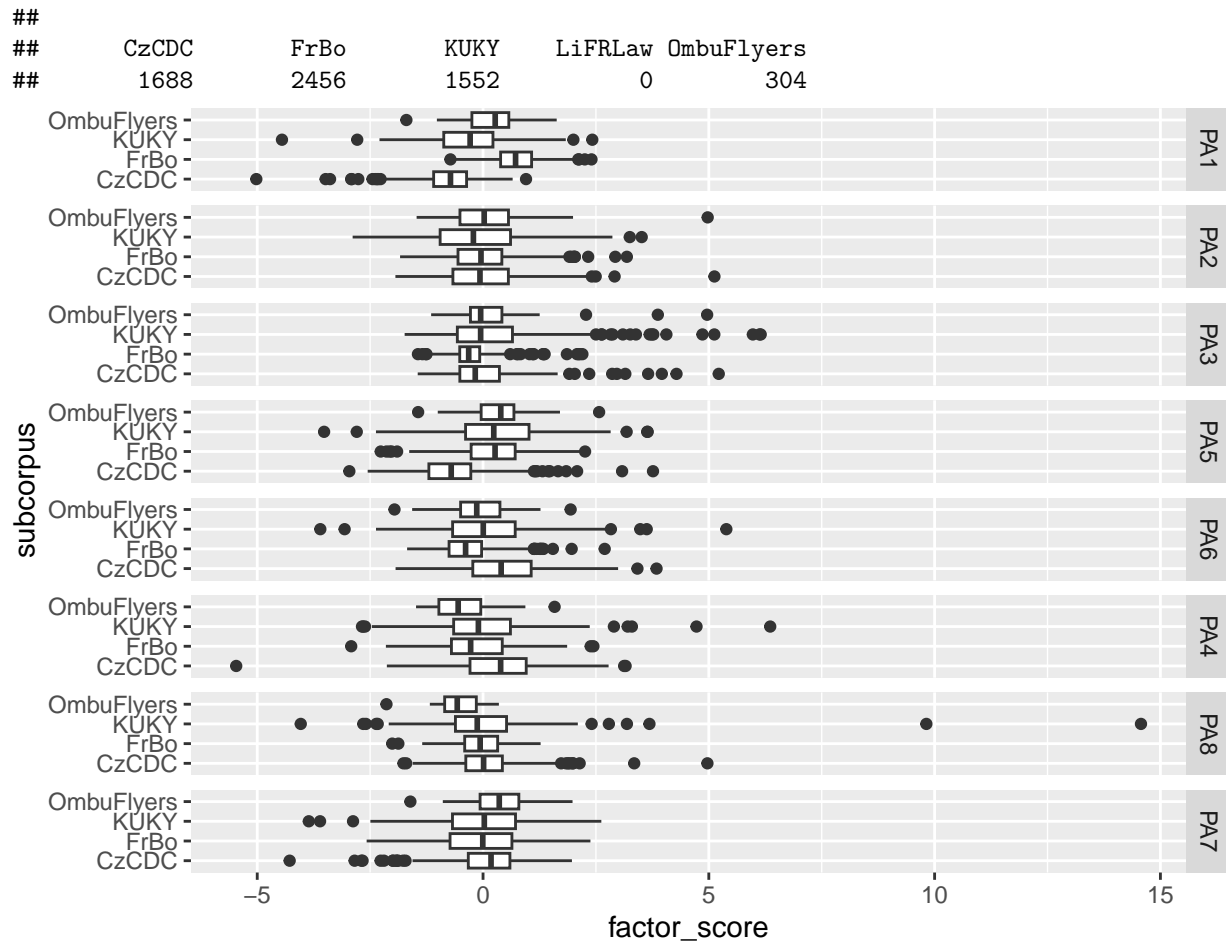
##          |      1.0000
##
##      KUKY |      1.858296      0.607432
##          |      0.6313      1.0000
##
##  LiFRLaw |      0.448053      0.226464      0.130070
##          |      1.0000      1.0000      1.0000
##
## OmbuFlye |      5.151007      4.527602      4.074951      1.079222
##          |      0.0000*      0.0001*      0.0005*      1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0359
##
## Test for the significance of differences in subcorpus over PA7 :
##
##      Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 9.544, df = 4, p-value = 0.05
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC      FrBo      KUKY      LiFRLaw
## -----+-----
##      FrBo |      1.582285
##          |      1.0000
##
##      KUKY |      0.671085     -0.814927
##          |      1.0000      1.0000
##
##  LiFRLaw |     -0.949506     -1.195481     -1.063657
##          |      1.0000      1.0000      1.0000
##
## OmbuFlye |     -1.842386     -2.710779     -2.206470      0.379197
##          |      0.6542      0.0671      0.2735      1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0127
##
##      factor kruskal_p epsilon2
## 1      PA1 2.19e-84 0.5260
## 2      PA2 2.09e-01 0.0078
## 3      PA3 1.69e-06 0.0429
## 4      PA5 2.60e-33 0.2110
## 5      PA6 2.20e-19 0.1250
## 6      PA4 1.13e-10 0.0697
## 7      PA8 2.02e-05 0.0359
## 8      PA7 4.89e-02 0.0127
##

```

```
## p < 5e-2 found in: PA1 PA3 PA5 PA6 PA4 PA8
## p < 1e-2 found in: PA1 PA3 PA5 PA6 PA4 PA8
## p < 1e-3 found in: PA1 PA3 PA5 PA6 PA4 PA8
## p < 1e-4 found in: PA1 PA3 PA5 PA6 PA4 PA8
```

subcorpus wo/ LiFRLaw

```
analyze_distributions(
  broad_data$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 = 395.0676, df = 3, p-value = 0
##
##
##           Comparison of x by group
##           (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC      FrBo      KUKY
```

```

## -----+-----
##      FrBo | -18.94981
##           | 0.0000*
##
##      KUKY | -5.093583 12.95203
##           | 0.0000* 0.0000*
##
## OmbuFlye | -5.882160 3.826214 -2.987223
##           | 0.0000* 0.0008* 0.0169*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.527
##
## Test for the significance of differences in subcorpus over PA2 :
##
##      Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 4.3463, df = 3, p-value = 0.23
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC      FrBo      KUKY
## -----+-----
##      FrBo | 0.037729
##           | 1.0000
##
##      KUKY | 1.596816 1.694989
##           | 0.6618 0.5405
##
## OmbuFlye | -0.629049 -0.664238 -1.520227
##           | 1.0000 1.0000 0.7707
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0058
##
## Test for the significance of differences in subcorpus over PA3 :
##
##      Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 28.8785, df = 3, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC      FrBo      KUKY
## -----+-----
##      FrBo | 3.410353

```



```

##          |      0.0039*
##          |
##      KUKY | -1.110008 -4.528926
##          |      1.0000      0.0000*
##          |
## OmbuFlye | -1.596167 -3.409067 -0.963214
##          |      0.6627      0.0039*      1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0386
##
## Test for the significance of differences in subcorpus over PA5 :
##
##      Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 154.436, df = 3, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC      FrBo      KUKY
## -----+-----
##      FrBo | -11.26250
##          |      0.0000*
##          |
##      KUKY | -9.949810      0.190224
##          |      0.0000*      1.0000
##          |
## OmbuFlye | -6.272823 -0.571540 -0.652392
##          |      0.0000*      1.0000      1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.206
##
## Test for the significance of differences in subcorpus over PA6 :
##
##      Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 92.2063, df = 3, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC      FrBo      KUKY
## -----+-----
##      FrBo |  9.524438
##          |      0.0000*
##          |

```

```

##      KUKY |    4.101429  -4.838276
##          |    0.0002*    0.0000*
##          |
## OmbuFlye |    3.102433  -1.773517    0.782204
##          |    0.0115*    0.4569    1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.123
##
## Test for the significance of differences in subcorpus over PA4 :
##
##      Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 52.3748, df = 3, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC      FrBo      KUKY
## -----+-----
##      FrBo |    6.339263
##          |    0.0000*
##          |
##      KUKY |    3.787601  -2.073076
##          |    0.0009*    0.2290
##          |
## OmbuFlye |    5.236072    2.069230    3.077682
##          |    0.0000*    0.2311    0.0125*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0699
##
## Test for the significance of differences in subcorpus over PA8 :
##
##      Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 26.8487, df = 3, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC      FrBo      KUKY
## -----+-----
##      FrBo |    1.445179
##          |    0.8904
##          |
##      KUKY |    1.855833    0.603630
##          |    0.3809    1.0000

```

```

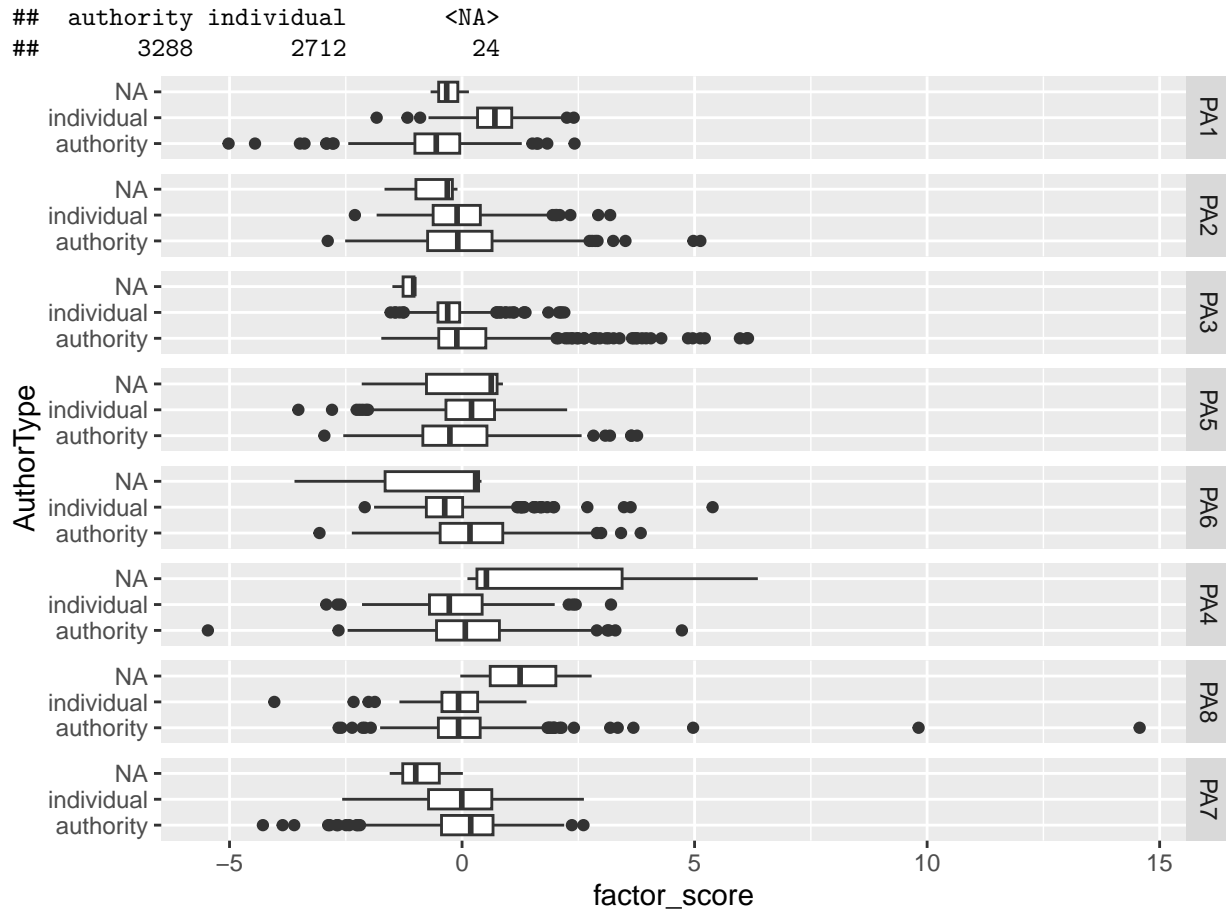
##      |
## OmbuFlye | 5.143837 4.519651 4.069209
##      | 0.0000* 0.0000* 0.0003*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0358
##
## Test for the significance of differences in subcorpus over PA7 :
##
## Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 8.4499, df = 3, p-value = 0.04
##
##
## Comparison of x by group
## (Bonferroni)
## Col Mean-|
## Row Mean | CzCDC FrBo KUKY
## -----+-----
## FrBo | 1.584972
##      | 0.6778
##
## KUKY | 0.674188 -0.814182
##      | 1.0000 1.0000
##
## OmbuFlye | -1.843864 -2.713691 -2.209678
##      | 0.3912 0.0399* 0.1628
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0113
##
## factor kruskal_p epsilon2
## 1 PA1 2.59e-85 0.5270
## 2 PA2 2.26e-01 0.0058
## 3 PA3 2.38e-06 0.0386
## 4 PA5 2.91e-33 0.2060
## 5 PA6 7.36e-20 0.1230
## 6 PA4 2.49e-11 0.0699
## 7 PA8 6.33e-06 0.0358
## 8 PA7 3.76e-02 0.0113
##
## p < 5e-2 found in: PA1 PA3 PA5 PA6 PA4 PA8 PA7
## p < 1e-2 found in: PA1 PA3 PA5 PA6 PA4 PA8
## p < 1e-3 found in: PA1 PA3 PA5 PA6 PA4 PA8
## p < 1e-4 found in: PA1 PA3 PA5 PA6 PA4 PA8

```

AuthorType

```
analyze_distributions(broad_data$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 = 355.7204, df = 1, p-value = 0
##
##
##           Comparison of x by group
##           (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## -----+-----
## individu | -18.86055
##           |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.473
##
## Test for the significance of differences in AuthorType over PA2 :
##
##   Kruskal-Wallis rank sum test
##
```

```

## data: x and group
## Kruskal-Wallis chi-squared = 0.4354, df = 1, p-value = 0.51
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## -----+-----
## individu |   0.659857
##          |   0.5093
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.000579
##
## Test for the significance of differences in AuthorType over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 27.246, df = 1, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## -----+-----
## individu |   5.219767
##          |   0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0362
##
## Test for the significance of differences in AuthorType over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 28.5227, df = 1, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## -----+-----
## individu |  -5.340665
##          |   0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha

```

```

## epsilon2 = 0.0379
##
## Test for the significance of differences in AuthorType over PA6 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 62.1615, df = 1, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## -----+-----
## individu |   7.884258
##          |   0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0827
##
## Test for the significance of differences in AuthorType over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 20.3275, df = 1, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## -----+-----
## individu |   4.508604
##          |   0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.027
##
## Test for the significance of differences in AuthorType over PA8 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.0469, df = 1, p-value = 0.83
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |   authorit

```

```

## -----+-----
## individu | -0.216566
##          | 0.8285
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 6.24e-05
##
## Test for the significance of differences in AuthorType over PA7 :
##
## Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 4.6003, df = 1, p-value = 0.03
##
##
## Comparison of x by group
## (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## -----+-----
## individu | 2.144833
##          | 0.0320*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.00612
##
## factor kruskal_p epsilon2
## 1 PA1 2.41e-79 4.73e-01
## 2 PA2 5.09e-01 5.79e-04
## 3 PA3 1.79e-07 3.62e-02
## 4 PA5 9.26e-08 3.79e-02
## 5 PA6 3.16e-15 8.27e-02
## 6 PA4 6.53e-06 2.70e-02
## 7 PA8 8.29e-01 6.24e-05
## 8 PA7 3.20e-02 6.12e-03
##
## p < 5e-2 found in: PA1 PA3 PA5 PA6 PA4 PA7
## p < 1e-2 found in: PA1 PA3 PA5 PA6 PA4
## p < 1e-3 found in: PA1 PA3 PA5 PA6 PA4
## p < 1e-4 found in: PA1 PA3 PA5 PA6 PA4

```

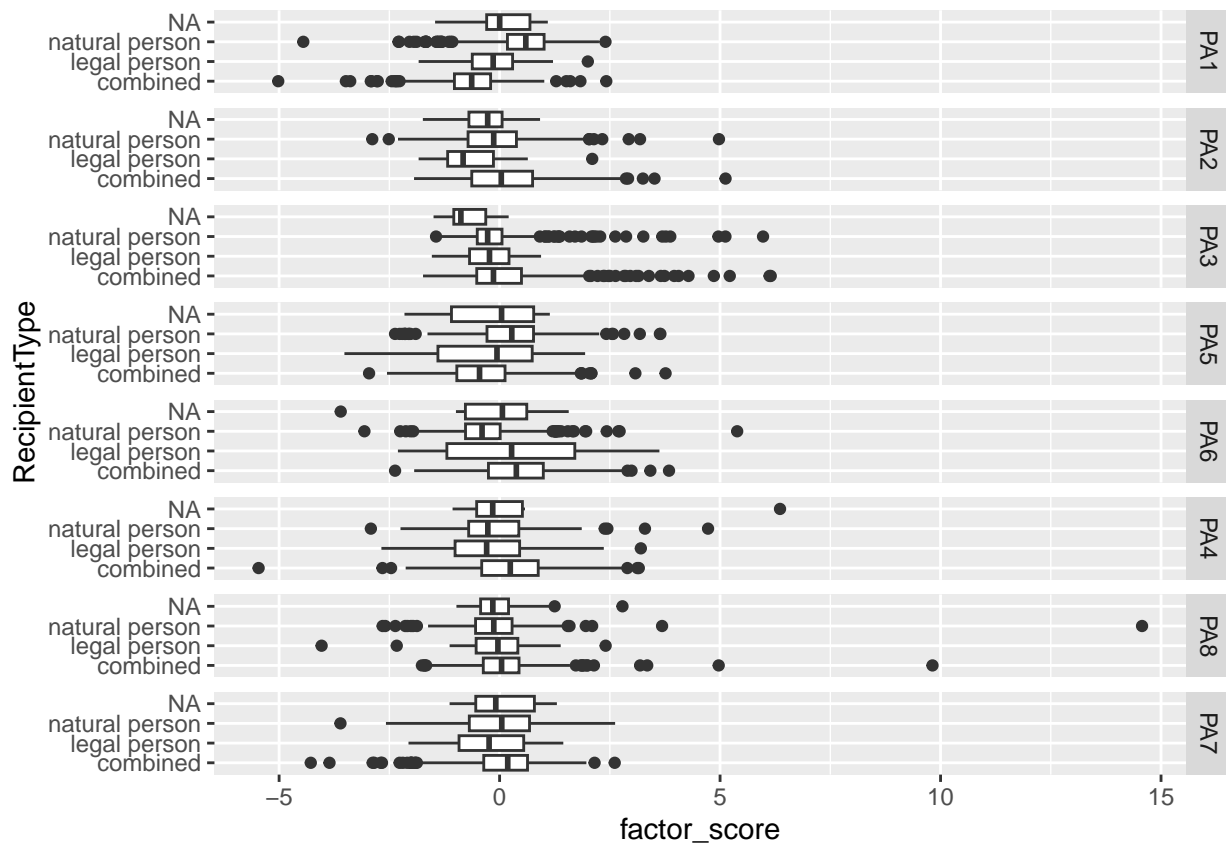
RecipientType

```
analyze_distributions(broad_data$long, "RecipientType")
```

```

##
## combined legal person natural person <NA>
## 2432 184 3304 104

```



```
##
## Test for the significance of differences in RecipientType over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 291.9381, df = 2, p-value = 0
##
##
##           Comparison of x by group
##           (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## -----+-----
## legal pe |  -2.490666
##           |    0.0383*
##           |
## natural  |  -17.05905  -3.503143
##           |    0.0000*    0.0014*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.388
##
## Test for the significance of differences in RecipientType over PA2 :
##
##   Kruskal-Wallis rank sum test
```



```

##
## data: x and group
## Kruskal-Wallis chi-squared = 18.4473, df = 2, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |    combined    legal pe
## -----+-----
## legal pe |    3.716618
##          |    0.0006*
##          |
## natural  |    2.856706  -2.743954
##          |    0.0128*    0.0182*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0245
##
## Test for the significance of differences in RecipientType over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 8.8831, df = 2, p-value = 0.01
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |    combined    legal pe
## -----+-----
## legal pe |    1.310350
##          |    0.5702
##          |
## natural  |    2.885911  -0.304734
##          |    0.0117*    1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0118
##
## Test for the significance of differences in RecipientType over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 85.8505, df = 2, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|

```

```

## Row Mean |      combined      legal pe
## -----+-----
## legal pe |    -0.677274
##          |         1.0000
##          |
## natural  |   -9.187961  -2.557227
##          |    0.0000*    0.0317*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.114
##
## Test for the significance of differences in RecipientType over PA6 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 106.4328, df = 2, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      combined      legal pe
## -----+-----
## legal pe |    1.168976
##          |         0.7272
##          |
## natural  |   10.27569   2.444572
##          |    0.0000*    0.0435*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.142
##
## Test for the significance of differences in RecipientType over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 34.625, df = 2, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      combined      legal pe
## -----+-----
## legal pe |    2.122712
##          |         0.1013
##          |
## natural  |    5.803025  -0.095787
##          |    0.0000*    1.0000
##

```

```

## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.046
##
## Test for the significance of differences in RecipientType over PA8 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 15.308, df = 2, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## -----+-----
## legal pe |   0.435293
##          |   1.0000
##          |
## natural  |   3.896288   0.934950
##          |   0.0003*    1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0204
##
## Test for the significance of differences in RecipientType over PA7 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 3.8949, df = 2, p-value = 0.14
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## -----+-----
## legal pe |   1.441846
##          |   0.4480
##          |
## natural  |   1.610202  -0.887450
##          |   0.3221    1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.00518
##
##   factor kruskal_p epsilon2
## 1   PA1  4.04e-64  0.38800
## 2   PA2  9.87e-05  0.02450
## 3   PA3  1.18e-02  0.01180

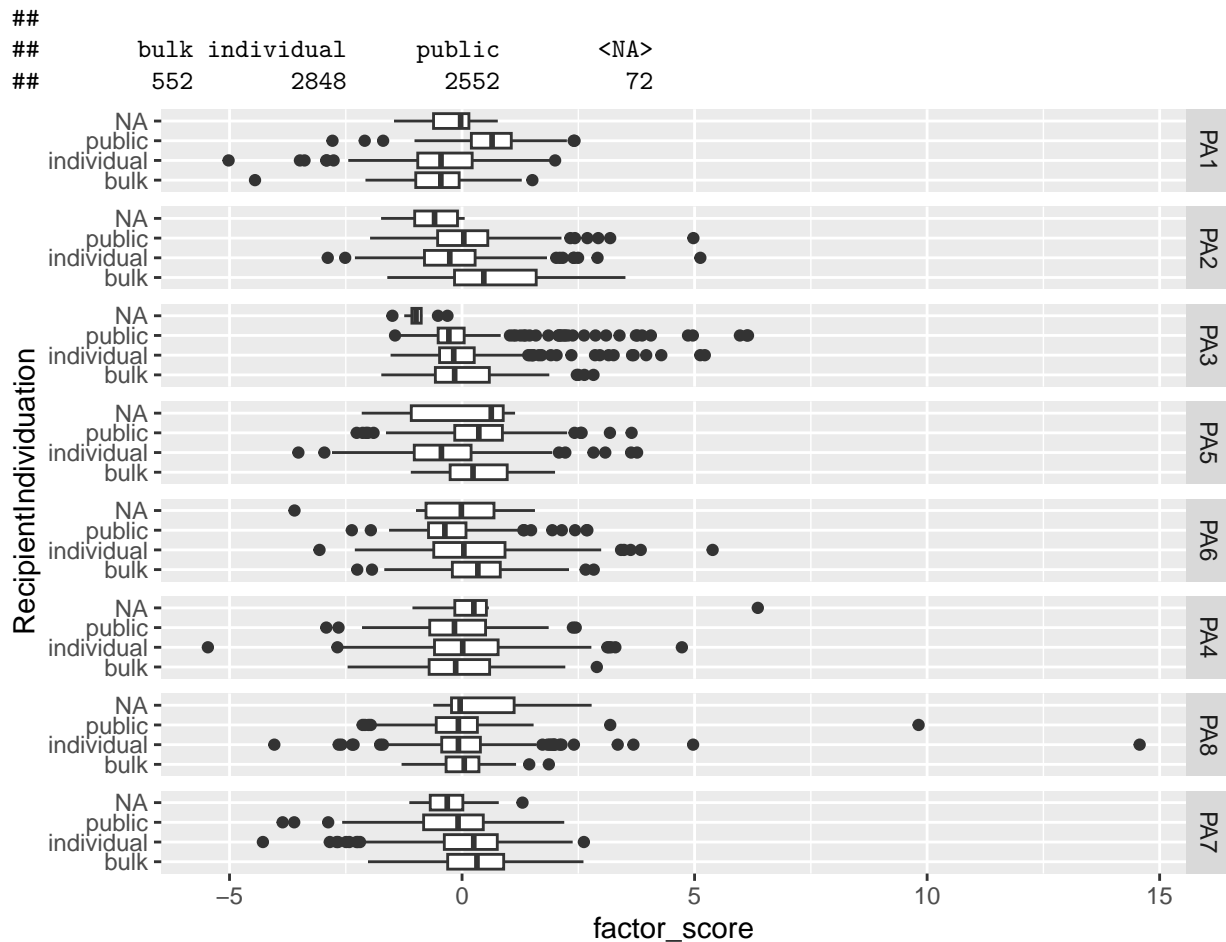
```

```
## 4    PA5  2.28e-19  0.11400
## 5    PA6  7.73e-24  0.14200
## 6    PA4  3.03e-08  0.04600
## 7    PA8  4.74e-04  0.02040
## 8    PA7  1.43e-01  0.00518
##
## p < 5e-2 found in: PA1 PA2 PA3 PA5 PA6 PA4 PA8
## p < 1e-2 found in: PA1 PA2 PA5 PA6 PA4 PA8
## p < 1e-3 found in: PA1 PA2 PA5 PA6 PA4 PA8
## p < 1e-4 found in: PA1 PA5 PA6 PA4
```

court decisions often with RecipientType = combined.

RecipientIndividuation

```
analyze_distributions(broad_data$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 = 231.7611, df = 2, p-value = 0
```

```

##
##
##           Comparison of x by group
##           (Bonferroni)
## Col Mean-|
## Row Mean |      bulk   individu
## -----+-----
## individu |   -0.802883
##           |       1.0000
##           |
##   public |   -9.148637  -14.38526
##           |       0.0000*   0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.308
##
## Test for the significance of differences in RecipientIndividuation over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 39.7178, df = 2, p-value = 0
##
##
##           Comparison of x by group
##           (Bonferroni)
## Col Mean-|
## Row Mean |      bulk   individu
## -----+-----
## individu |    5.819968
##           |    0.0000*
##           |
##   public |    3.480791   -3.935297
##           |    0.0015*   0.0002*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0528
##
## Test for the significance of differences in RecipientIndividuation over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 6.1779, df = 2, p-value = 0.05
##
##
##           Comparison of x by group
##           (Bonferroni)
## Col Mean-|
## Row Mean |      bulk   individu
## -----+-----
## individu |    0.583560

```

```

##          |      1.0000
##          |
## public | 1.832342 2.159889
##          |      0.2007      0.0923
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.00822
##
## Test for the significance of differences in RecipientIndividuation over PA5 :
##
## Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 117.9317, df = 2, p-value = 0
##
##
## Comparison of x by group
## (Bonferroni)
## Col Mean-|
## Row Mean |      bulk      individu
## -----+-----
## individu | 5.787178
##          | 0.0000*
##          |
## public | -0.324537 -10.43260
##          | 1.0000      0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.157
##
## Test for the significance of differences in RecipientIndividuation over PA6 :
##
## Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 46.2243, df = 2, p-value = 0
##
##
## Comparison of x by group
## (Bonferroni)
## Col Mean-|
## Row Mean |      bulk      individu
## -----+-----
## individu | 1.848811
##          | 0.1935
##          |
## public | 5.184785 5.774522
##          | 0.0000*      0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0615

```

```

##
## Test for the significance of differences in RecipientIndividuation over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 5.8732, df = 2, p-value = 0.05
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      bulk   individu
## -----+-----
## individu |  -0.767062
##           |      1.0000
##           |
##   public |   0.646113   2.421398
##           |      1.0000   0.0464*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.00781
##
## Test for the significance of differences in RecipientIndividuation over PA8 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 3.3278, df = 2, p-value = 0.19
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      bulk   individu
## -----+-----
## individu |   0.967678
##           |      0.9996
##           |
##   public |   1.665889   1.217876
##           |      0.2872   0.6698
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.00443
##
## Test for the significance of differences in RecipientIndividuation over PA7 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 25.8542, df = 2, p-value = 0
##

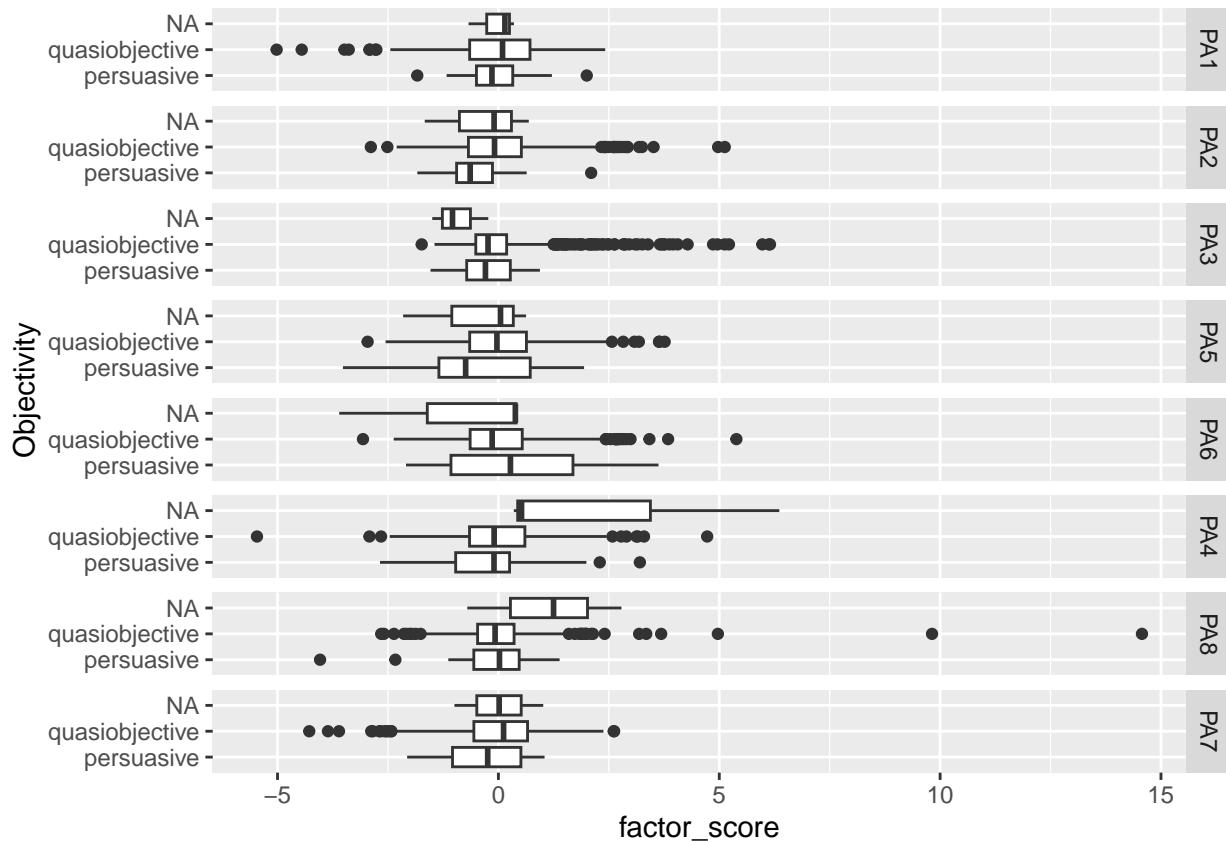
```

```
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      bulk   individu
## -----+-----
## individu |    1.186493
##          |    0.7063
##          |
##   public |    3.743362   4.422216
##          |    0.0005*   0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0344
##
##   factor kruskal_p epsilon2
## 1   PA1  4.72e-51  0.30800
## 2   PA2  2.37e-09  0.05280
## 3   PA3  4.55e-02  0.00822
## 4   PA5  2.46e-26  0.15700
## 5   PA6  9.17e-11  0.06150
## 6   PA4  5.30e-02  0.00781
## 7   PA8  1.89e-01  0.00443
## 8   PA7  2.43e-06  0.03440
##
## p < 5e-2 found in: PA1 PA2 PA5 PA6 PA4 PA7
## p < 1e-2 found in: PA1 PA2 PA5 PA6 PA7
## p < 1e-3 found in: PA1 PA2 PA5 PA6 PA7
## p < 1e-4 found in: PA1 PA2 PA5 PA6 PA7
```

Objectivity

```
analyze_distributions(broad_data$long, "Objectivity")
```

```
##
##   persuasive quasiobjective      <NA>
##         168             5832        24
```

```
##
## Test for the significance of differences in Objectivity over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.2128, df = 1, p-value = 0.64
##
##
##               Comparison of x by group
##               (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## -----+-----
## quasiobj | -0.461269
##          |    0.6446
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.000283
##
## Test for the significance of differences in Objectivity over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 5.7127, df = 1, p-value = 0.02
```

```

##
##
##           Comparison of x by group
##           (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## -----+-----
## quasiobj |  -2.390123
##           |    0.0168*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0076
##
## Test for the significance of differences in Objectivity over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.7303, df = 1, p-value = 0.39
##
##
##           Comparison of x by group
##           (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## -----+-----
## quasiobj |  -0.854600
##           |    0.3928
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.000971
##
## Test for the significance of differences in Objectivity over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 2.7409, df = 1, p-value = 0.1
##
##
##           Comparison of x by group
##           (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## -----+-----
## quasiobj |  -1.655565
##           |    0.0978
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.00364
##

```

```

## Test for the significance of differences in Objectivity over PA6 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.7585, df = 1, p-value = 0.38
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## -----+-----
## quasiobj |   0.870946
##           |   0.3838
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.00101
##
## Test for the significance of differences in Objectivity over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.7044, df = 1, p-value = 0.4
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## -----+-----
## quasiobj | -0.839276
##           |   0.4013
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.000937
##
## Test for the significance of differences in Objectivity over PA8 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.314, df = 1, p-value = 0.58
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## -----+-----
## quasiobj |   0.560368

```

```

##          |      0.5752
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.000418
##
## Test for the significance of differences in Objectivity over PA7 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 2.8021, df = 1, p-value = 0.09
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## -----+-----
## quasiobj | -1.673954
##          |      0.0941
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.00373
##
##   factor kruskal_p epsilon2
## 1    PA1    0.6446 0.000283
## 2    PA2    0.0168 0.007600
## 3    PA3    0.3928 0.000971
## 4    PA5    0.0978 0.003640
## 5    PA6    0.3838 0.001010
## 6    PA4    0.4013 0.000937
## 7    PA8    0.5752 0.000418
## 8    PA7    0.0941 0.003730
##
## p < 5e-2 found in: PA2
## p < 1e-2 found in:
## p < 1e-3 found in:
## p < 1e-4 found in:

```

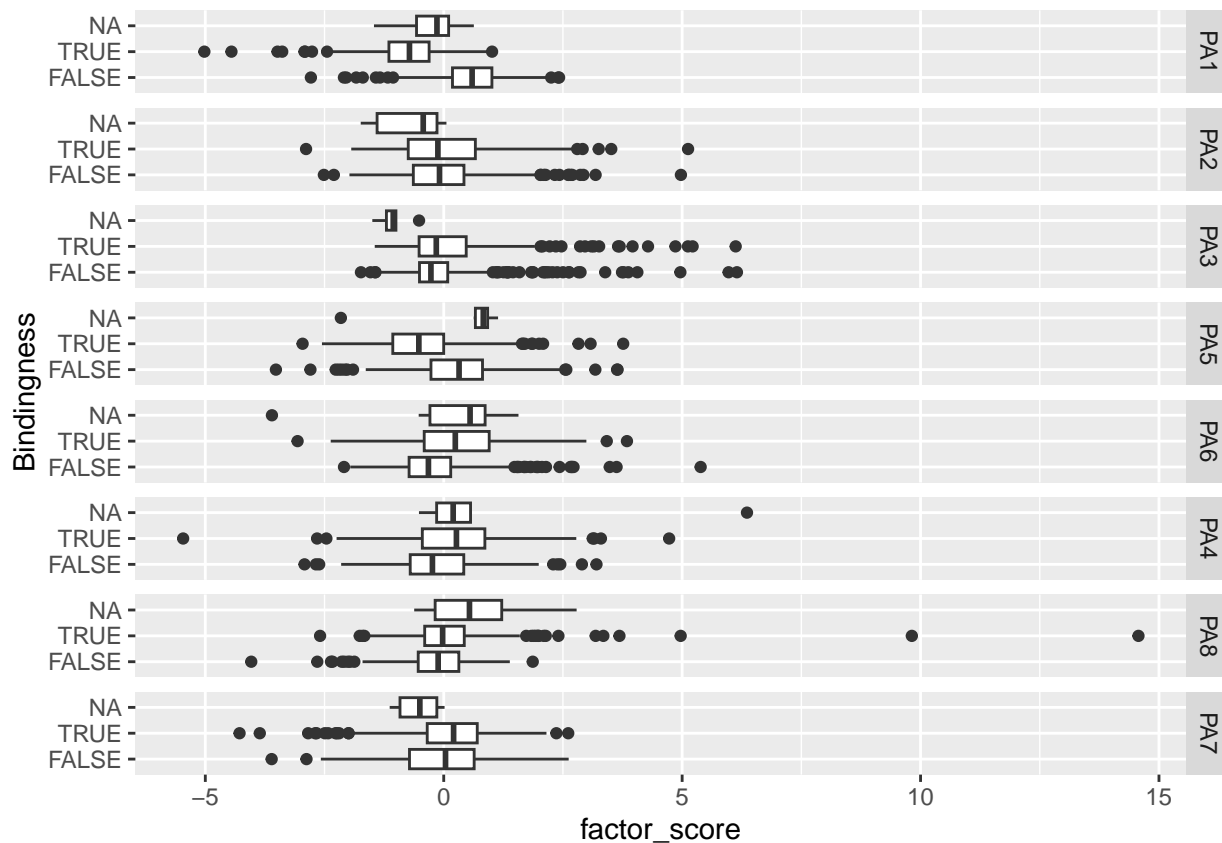
Bindingness

```
analyze_distributions(broad_data$long, "Bindingness")
```

```

##
## FALSE  TRUE  <NA>
## 3552  2424    48

```



```
##
## Test for the significance of differences in Bindingness over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 380.9685, df = 1, p-value = 0
##
##
##               Comparison of x by group
##               (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## -----+-----
##   TRUE |   19.51841
##       |   0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.507
##
## Test for the significance of differences in Bindingness over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.0529, df = 1, p-value = 0.82
```

```

##
##
##           Comparison of x by group
##           (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## -----+-----
##      TRUE |  -0.229985
##           |    0.8181
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 7.03e-05
##
## Test for the significance of differences in Bindingness over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 7.2737, df = 1, p-value = 0.01
##
##
##           Comparison of x by group
##           (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## -----+-----
##      TRUE |  -2.696982
##           |    0.0070*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.00967
##
## Test for the significance of differences in Bindingness over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 118.7006, df = 1, p-value = 0
##
##
##           Comparison of x by group
##           (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## -----+-----
##      TRUE |   10.89497
##           |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.158
##

```

```

## Test for the significance of differences in Bindingness over PA6 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 49.5439, df = 1, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## -----+-----
##      TRUE |  -7.038743
##           |    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 = 30.6385, df = 1, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## -----+-----
##      TRUE |  -5.535201
##           |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0407
##
## Test for the significance of differences in Bindingness over PA8 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 9.501, df = 1, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## -----+-----
##      TRUE |  -3.082363

```

```
##          |    0.0021*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0126
##
## Test for the significance of differences in Bindingness over PA7 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 5.5352, df = 1, p-value = 0.02
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |    FALSE
## -----+-----
##   TRUE | -2.352693
##        |    0.0186*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.00736
##
##   factor kruskal_p epsilon2
## 1   PA1  7.66e-85 5.07e-01
## 2   PA2  8.18e-01 7.03e-05
## 3   PA3  7.00e-03 9.67e-03
## 4   PA5  1.22e-27 1.58e-01
## 5   PA6  1.94e-12 6.59e-02
## 6   PA4  3.11e-08 4.07e-02
## 7   PA8  2.05e-03 1.26e-02
## 8   PA7  1.86e-02 7.36e-03
##
## p < 5e-2 found in: PA1 PA3 PA5 PA6 PA4 PA8 PA7
## p < 1e-2 found in: PA1 PA3 PA5 PA6 PA4 PA8
## p < 1e-3 found in: PA1 PA5 PA6 PA4
## p < 1e-4 found in: PA1 PA5 PA6 PA4
```

Feature-factor correlations

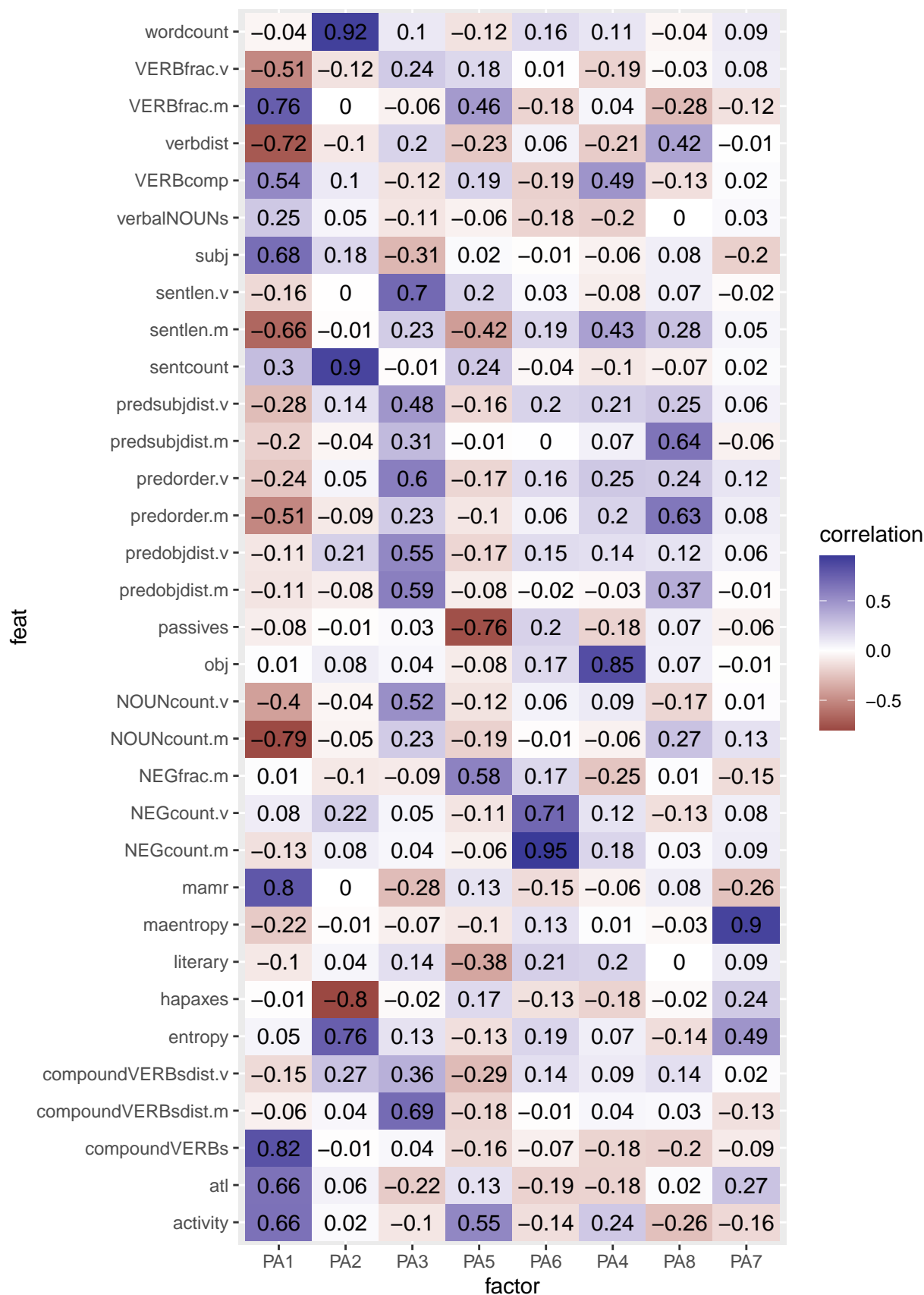
```
broad_data_factors_corr <- broad_data$feat_long %>%
  group_by(feat, factor) %>%
  summarize(correlation = cor(feat_value, factor_score))
```

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

```
broad_data_factors_corr %>%
  filter(feat %in% rownames(fa_broad$loadings)) %>%
  ggplot(aes(
    x = factor,
```



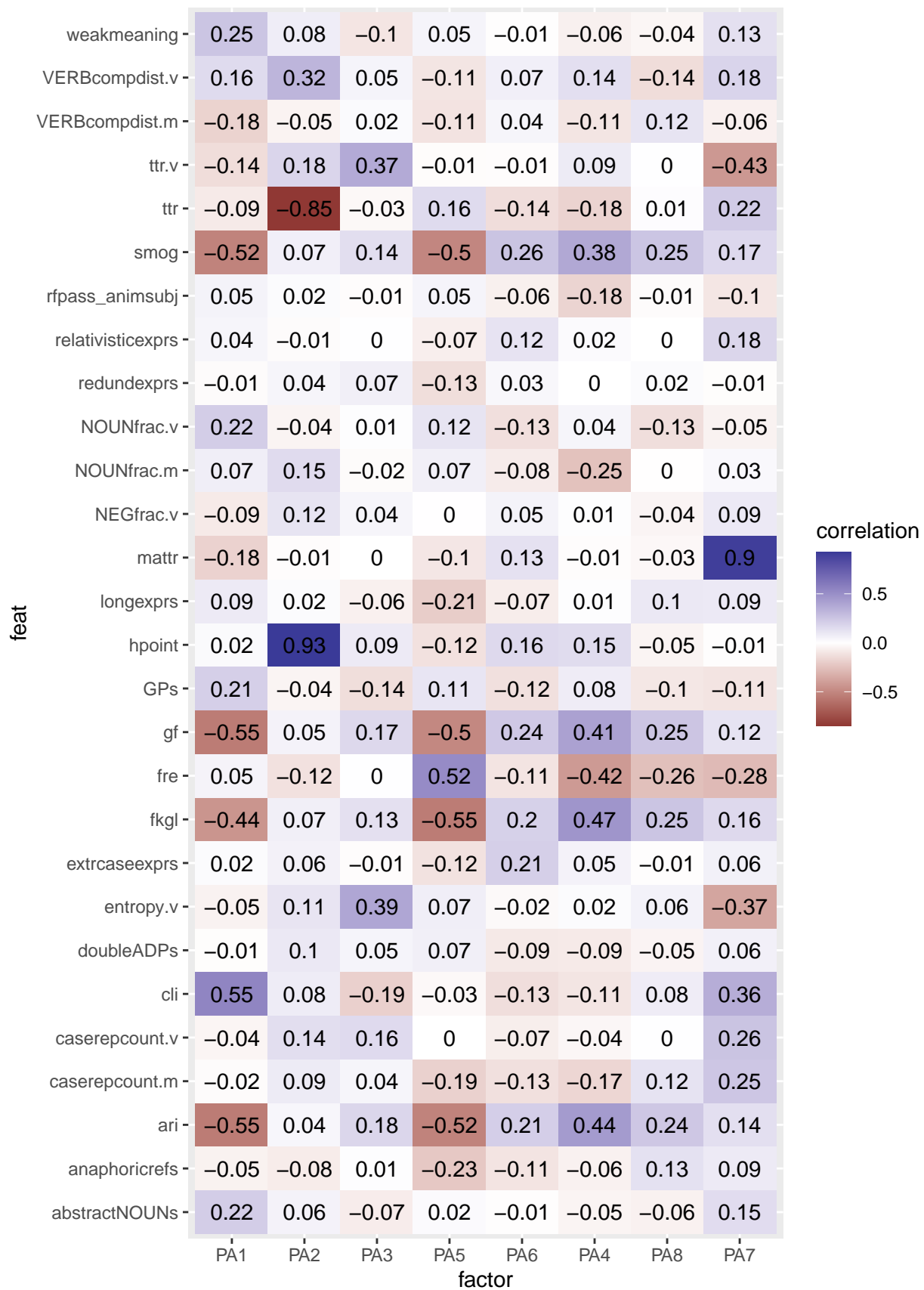
```
y = feat,  
  fill = correlation,  
  label = round(correlation, 2)  
) +  
geom_tile() +  
geom_text() +  
scale_fill_gradient2()
```



```

broad_data_factors_corr %>%
  filter(!(feat %in% rownames(fa_broad$loadings))) %>%
  ggplot(aes(
    x = factor,
    y = feat,
    fill = correlation,
    label = round(correlation, 2)
  )) +
  geom_tile() +
  geom_text() +
  scale_fill_gradient2()

```

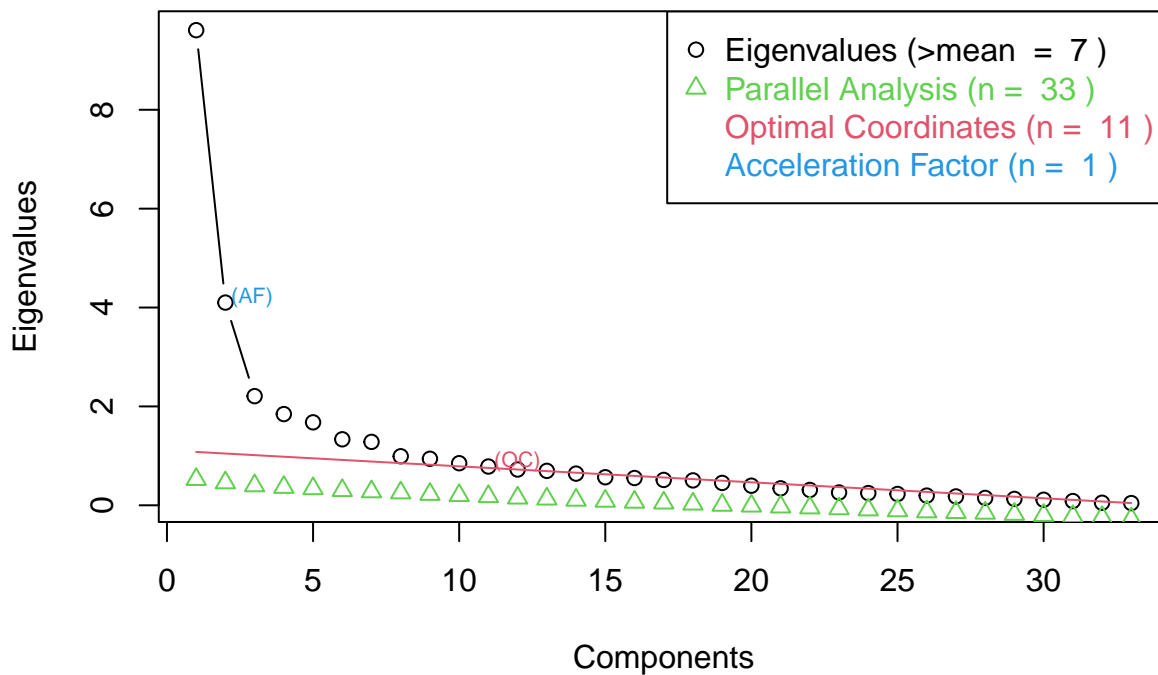


first FA

No. of factors

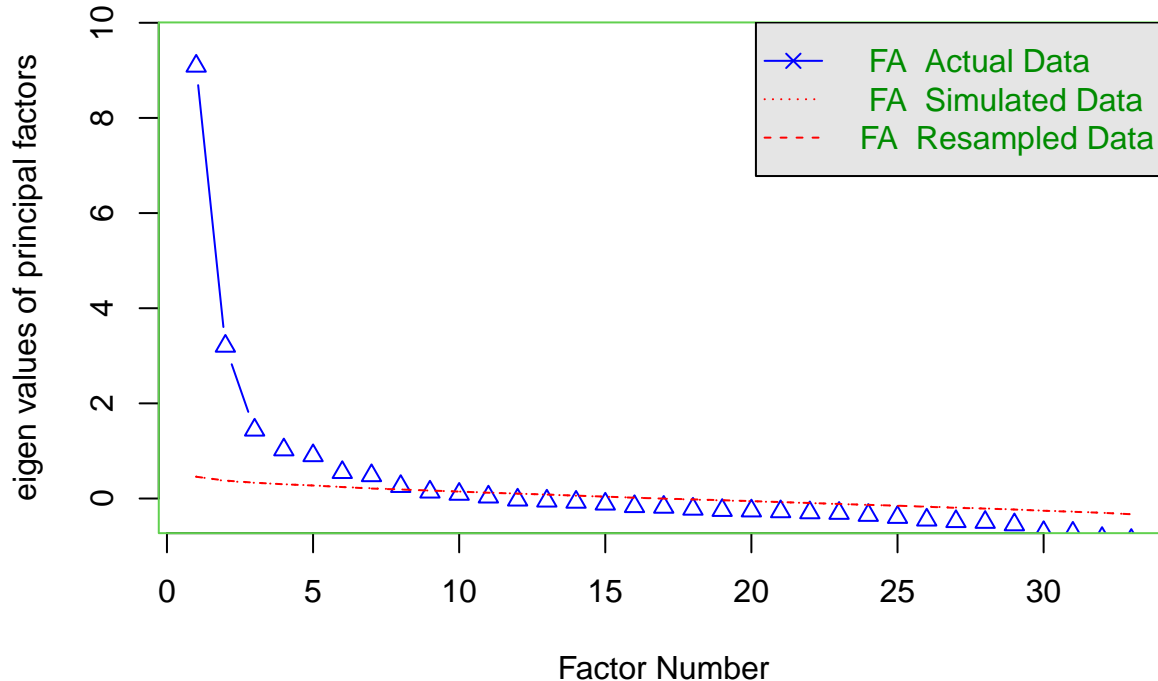
```
eigen <- eigen(cor(data_scaled))
par <- nFactors::parallel(
  subject = nrow(data_scaled),
  var = ncol(data_scaled),
  rep = 100,
  quantile = .95,
  model = "factors"
)
scree <- nScree(x = eigen$values, aparallel = par$eigen$gevpea)
plotnScree(scree)
```

Non Graphical Solutions to Scree Test



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

Parallel Analysis Scree Plots



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>

```
set.seed(42)
```

```
fa_1 <- fa(
  data_scaled,
  nfactors = 8,
  fm = "pa",
  rotate = "promax",
  oblique.scores = TRUE,
  scores = "tenBerge",
  n.iter = 100
)
fa_1
```

```
## Factor Analysis with confidence intervals using method = fa(r = data_scaled, nfactors = 8, n.iter = 100,
##   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	PA3	PA5	PA6	PA4	PA8	PA7	h2	u2
## sentlen.m	-0.62	-0.02	-0.03	-0.28	0.00	0.37	0.15	-0.02	0.94	0.063
## sentcount	0.15	0.96	0.03	0.32	-0.07	-0.16	0.00	-0.01	0.93	0.066
## atl	0.70	0.00	-0.02	0.06	-0.05	-0.13	0.10	0.30	0.57	0.431
## activity	0.66	-0.01	0.10	0.47	0.00	0.31	-0.09	-0.09	0.89	0.106
## VERBfrac.m	0.80	-0.06	0.20	0.35	-0.02	0.10	-0.12	-0.05	0.90	0.100

## wordcount	-0.15	0.95	0.00	0.01	0.02	0.00	-0.05	0.01	0.89	0.114
## entropy	0.03	0.72	0.07	-0.02	0.10	-0.04	-0.12	0.39	0.86	0.141
## sentlen.v	0.00	-0.01	0.73	0.28	0.01	-0.15	0.05	-0.02	0.46	0.538
## predsubjdist.m	-0.08	-0.04	0.25	0.12	-0.04	0.06	0.55	-0.04	0.45	0.555
## compoundVERBs	0.99	-0.15	0.30	-0.31	0.07	-0.18	-0.14	-0.04	0.70	0.298
## passives	0.03	-0.09	-0.03	-0.79	0.15	-0.25	-0.06	-0.09	0.57	0.427
## predobjdist.m	0.08	-0.12	0.60	0.01	-0.05	-0.08	0.29	0.00	0.42	0.583
## literary	0.00	-0.04	0.07	-0.34	0.15	0.14	-0.05	0.06	0.24	0.758
## verbdist	-0.74	0.00	0.00	-0.12	-0.06	-0.25	0.26	-0.04	0.81	0.188
## maentropy	-0.19	-0.07	-0.15	-0.03	0.12	-0.01	-0.01	0.82	0.76	0.245
## predorder.m	-0.45	-0.07	0.06	0.06	-0.04	0.19	0.51	0.07	0.70	0.297
## hapaxes	0.10	-0.83	0.07	0.07	0.01	-0.10	0.01	0.29	0.72	0.282
## VERBcomp	0.56	0.02	-0.01	0.15	-0.15	0.54	-0.01	0.04	0.60	0.404
## NOUNcount.v	-0.33	-0.04	0.43	-0.08	-0.05	0.01	-0.22	-0.03	0.41	0.594
## subj	0.69	0.12	-0.14	-0.04	0.11	-0.02	0.13	-0.14	0.58	0.422
## NOUNcount.m	-0.84	0.05	0.01	-0.08	-0.17	-0.10	0.14	0.07	0.79	0.209
## predobjdist.v	0.05	0.14	0.51	-0.07	0.07	0.04	0.07	0.02	0.39	0.606
## NEGcount.m	0.04	-0.05	-0.06	0.08	1.00	0.08	0.03	0.09	0.94	0.063
## compoundVERBsdist.m	0.13	-0.02	0.71	-0.14	-0.08	-0.04	-0.03	-0.14	0.43	0.566
## VERBfrac.v	-0.55	-0.03	0.15	0.23	-0.04	-0.21	-0.06	0.06	0.35	0.648
## NEGcount.v	0.21	0.09	0.01	-0.03	0.75	0.02	-0.11	0.07	0.59	0.415
## compoundVERBsdist.v	-0.07	0.23	0.28	-0.20	0.04	0.00	0.06	-0.03	0.33	0.672
## predsubjdist.v	-0.14	0.10	0.38	-0.03	0.10	0.13	0.17	0.03	0.47	0.533
## mamr	0.84	-0.07	-0.06	0.02	0.01	0.02	0.16	-0.17	0.77	0.234
## obj	0.08	-0.03	-0.06	0.00	0.08	0.83	0.10	-0.02	0.68	0.322
## predorder.v	-0.05	-0.02	0.52	-0.05	0.07	0.16	0.17	0.08	0.54	0.463
## verbalNOUNs	0.23	0.05	-0.02	-0.12	-0.14	-0.18	0.00	0.04	0.14	0.862
## NEGfrac.m	-0.03	-0.02	-0.03	0.60	0.29	-0.21	0.09	-0.09	0.40	0.602
##	com									
## sentlen.m	2.2									
## sentcount	1.3									
## atl	1.5									
## activity	2.4									
## VERBfrac.m	1.6									
## wordcount	1.1									
## entropy	1.7									
## sentlen.v	1.4									
## predsubjdist.m	1.6									
## compoundVERBs	1.6									
## passives	1.4									
## predobjdist.m	1.6									
## literary	2.0									
## verbdist	1.6									
## maentropy	1.2									
## predorder.m	2.4									
## hapaxes	1.3									
## VERBcomp	2.3									
## NOUNcount.v	2.6									
## subj	1.4									
## NOUNcount.m	1.2									
## predobjdist.v	1.3									
## NEGcount.m	1.1									
## compoundVERBsdist.m	1.3									
## VERBfrac.v	1.9									

```

## NEGcount.v          1.3
## compoundVERBsdist.v 3.1
## predsubjdist.v     2.4
## mamr                1.2
## obj                 1.1
## predorder.v        1.6
## verbalNOUNs        3.4
## NEGfrac.m          1.9
##
##
##          PA1  PA2  PA3  PA5  PA6  PA4  PA8  PA7
## SS loadings      6.71 3.10 2.53 2.08 1.74 1.56 1.29 1.19
## Proportion Var    0.20 0.09 0.08 0.06 0.05 0.05 0.04 0.04
## Cumulative Var    0.20 0.30 0.37 0.44 0.49 0.54 0.58 0.61
## Proportion Explained 0.33 0.15 0.13 0.10 0.09 0.08 0.06 0.06
## Cumulative Proportion 0.33 0.49 0.61 0.71 0.80 0.88 0.94 1.00
##
## With factor correlations of
##          PA1  PA2  PA3  PA5  PA6  PA4  PA8  PA7
## PA1  1.00  0.11 -0.56  0.38 -0.37 -0.18 -0.36 -0.17
## PA2  0.11  1.00  0.17 -0.26  0.27  0.25  0.01  0.18
## PA3 -0.56  0.17  1.00 -0.33  0.30  0.32  0.24  0.11
## PA5  0.38 -0.26 -0.33  1.00 -0.34 -0.23 -0.38 -0.17
## PA6 -0.37  0.27  0.30 -0.34  1.00  0.32  0.11  0.07
## PA4 -0.18  0.25  0.32 -0.23  0.32  1.00  0.00  0.08
## PA8 -0.36  0.01  0.24 -0.38  0.11  0.00  1.00 -0.10
## PA7 -0.17  0.18  0.11 -0.17  0.07  0.08 -0.10  1.00
##
## Mean item complexity = 1.7
## Test of the hypothesis that 8 factors are sufficient.
##
## df null model = 528 with the objective function = 24.21 with Chi Square = 17922.49
## df of the model are 292 and the objective function was 2.94
##
## 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 514.88 with prob < 1.6e-14
## The total n.obs was 753 with Likelihood Chi Square = 2157.52 with prob < 2.7e-281
##
## Tucker Lewis Index of factoring reliability = 0.805
## RMSEA index = 0.092 and the 90 % confidence intervals are 0.089 0.096
## BIC = 223.3
## Fit based upon off diagonal values = 0.99
## Measures of factor score adequacy
##
##          PA1  PA2  PA3  PA5  PA6  PA4
## Correlation of (regression) scores with factors 0.98 0.98 0.92 0.94 0.98 0.94
## Multiple R square of scores with factors 0.96 0.96 0.85 0.89 0.96 0.89
## Minimum correlation of possible factor scores 0.92 0.92 0.70 0.77 0.91 0.78
##
##          PA8  PA7
## Correlation of (regression) scores with factors 0.87 0.91
## Multiple R square of scores with factors 0.75 0.82
## Minimum correlation of possible factor scores 0.50 0.65
##
## Coefficients and bootstrapped confidence intervals

```


##		low	PA1	upper	low	PA2	upper	low	PA3	upper	low
## sentlen.m		-0.79	-0.62	-0.44	-0.06	-0.02	0.00	-0.09	-0.03	0.05	-0.33
## sentcount		0.10	0.15	0.21	0.91	0.96	1.01	-0.01	0.03	0.07	0.24
## atl		0.48	0.70	0.82	-0.05	0.00	0.07	-0.11	-0.02	0.07	-0.07
## activity		0.46	0.66	0.89	-0.04	-0.01	0.02	0.03	0.10	0.15	0.38
## VERBfrac.m		0.55	0.80	1.08	-0.09	-0.06	-0.01	0.09	0.20	0.31	0.26
## wordcount		-0.19	-0.15	-0.08	0.91	0.95	0.98	-0.04	0.00	0.04	-0.02
## entropy		-0.03	0.03	0.06	0.68	0.72	0.76	0.02	0.07	0.11	-0.07
## sentlen.v		-0.10	0.00	0.07	-0.06	-0.01	0.06	0.49	0.73	1.00	0.18
## predsubjdist.m		-0.33	-0.08	0.04	-0.09	-0.04	0.02	0.12	0.25	0.44	-0.05
## compoundVERBs		0.71	0.99	1.29	-0.21	-0.15	-0.09	0.18	0.30	0.41	-0.40
## passives		-0.04	0.03	0.11	-0.14	-0.09	-0.05	-0.10	-0.03	0.04	-0.89
## predobjdist.m		-0.08	0.08	0.19	-0.18	-0.12	-0.05	0.37	0.60	0.91	-0.15
## literary		-0.10	0.00	0.11	-0.10	-0.04	0.02	-0.05	0.07	0.19	-0.41
## verbdist		-1.00	-0.74	-0.55	-0.04	0.00	0.02	-0.04	0.00	0.07	-0.29
## maentropy		-0.35	-0.19	-0.12	-0.10	-0.07	-0.01	-0.23	-0.15	-0.09	-0.14
## predorder.m		-0.81	-0.45	-0.25	-0.11	-0.07	-0.01	-0.03	0.06	0.20	-0.14
## hapaxes		0.00	0.10	0.16	-0.86	-0.83	-0.78	-0.01	0.07	0.13	0.00
## VERBcomp		0.39	0.56	0.74	-0.03	0.02	0.07	-0.08	-0.01	0.06	0.06
## NOUNcount.v		-0.43	-0.33	-0.15	-0.11	-0.04	0.03	0.25	0.43	0.62	-0.13
## subj		0.50	0.69	0.82	0.07	0.12	0.18	-0.23	-0.14	-0.07	-0.14
## NOUNcount.m		-1.12	-0.84	-0.59	0.00	0.05	0.09	-0.06	0.01	0.09	-0.15
## predobjdist.v		-0.10	0.05	0.19	0.05	0.14	0.24	0.29	0.51	0.79	-0.19
## NEGcount.m		-0.06	0.04	0.09	-0.08	-0.05	-0.01	-0.13	-0.06	-0.01	-0.02
## compoundVERBsdist.m		0.01	0.13	0.27	-0.09	-0.02	0.06	0.50	0.71	0.96	-0.22
## VERBfrac.v		-0.73	-0.55	-0.36	-0.10	-0.03	0.05	0.03	0.15	0.28	0.12
## NEGcount.v		0.14	0.21	0.31	0.04	0.09	0.13	-0.05	0.01	0.08	-0.11
## compoundVERBsdist.v		-0.20	-0.07	0.04	0.16	0.23	0.31	0.10	0.28	0.48	-0.31
## predsubjdist.v		-0.32	-0.14	-0.02	0.05	0.10	0.17	0.19	0.38	0.60	-0.15
## mamr		0.60	0.84	1.02	-0.11	-0.07	-0.01	-0.13	-0.06	0.02	-0.08
## obj		0.01	0.08	0.14	-0.07	-0.03	0.01	-0.14	-0.06	0.02	-0.05
## predorder.v		-0.22	-0.05	0.06	-0.08	-0.02	0.06	0.29	0.52	0.80	-0.17
## verbalNOUNs		0.10	0.23	0.34	-0.03	0.05	0.13	-0.14	-0.02	0.08	-0.23
## NEGfrac.m		-0.16	-0.03	0.05	-0.08	-0.02	0.04	-0.11	-0.03	0.05	0.46
##		PA5	upper	low	PA6	upper	low	PA4	upper	low	PA8
## sentlen.m		-0.28	-0.22	-0.03	0.00	0.05	0.24	0.37	0.53	-0.22	0.15
## sentcount		0.32	0.36	-0.13	-0.07	-0.02	-0.23	-0.16	-0.12	-0.30	0.00
## atl		0.06	0.11	-0.19	-0.05	0.06	-0.24	-0.13	-0.05	-0.59	0.10
## activity		0.47	0.57	-0.05	0.00	0.05	0.22	0.31	0.43	-0.29	-0.09
## VERBfrac.m		0.35	0.44	-0.06	-0.02	0.03	0.05	0.10	0.16	-0.55	-0.12
## wordcount		0.01	0.06	-0.01	0.02	0.07	-0.04	0.00	0.03	-0.14	-0.05
## entropy		-0.02	0.02	0.06	0.10	0.15	-0.09	-0.04	0.00	-0.65	-0.12
## sentlen.v		0.28	0.36	-0.07	0.01	0.07	-0.24	-0.15	-0.05	-0.11	0.05
## predsubjdist.m		0.12	0.22	-0.18	-0.04	0.08	-0.06	0.06	0.23	0.05	0.55
## compoundVERBs		-0.31	-0.21	0.02	0.07	0.14	-0.26	-0.18	-0.11	-0.67	-0.14
## passives		-0.79	-0.63	0.08	0.15	0.24	-0.39	-0.25	-0.14	-0.16	-0.06
## predobjdist.m		0.01	0.11	-0.18	-0.05	0.03	-0.15	-0.08	0.01	-0.27	0.29
## literary		-0.34	-0.23	0.06	0.15	0.28	0.05	0.14	0.25	-0.17	-0.05
## verbdist		-0.12	-0.02	-0.13	-0.06	-0.01	-0.35	-0.25	-0.17	0.01	0.26
## maentropy		-0.03	0.02	0.06	0.12	0.18	-0.07	-0.01	0.04	-0.79	-0.01
## predorder.m		0.06	0.14	-0.17	-0.04	0.08	0.04	0.19	0.32	0.04	0.51
## hapaxes		0.07	0.11	-0.06	0.01	0.06	-0.16	-0.10	-0.05	-0.27	0.01
## VERBcomp		0.15	0.22	-0.23	-0.15	-0.08	0.36	0.54	0.80	-0.21	-0.01
## NOUNcount.v		-0.08	0.07	-0.14	-0.05	0.08	-0.08	0.01	0.10	-0.43	-0.22

## subj	-0.04	0.01	0.01	0.11	0.17	-0.09	-0.02	0.05	-0.16	0.13
## NOUNcount.m	-0.08	-0.02	-0.27	-0.17	-0.09	-0.18	-0.10	-0.03	-0.14	0.14
## predobjdist.v	-0.07	0.04	-0.04	0.07	0.16	-0.04	0.04	0.14	-0.22	0.07
## NEGcount.m	0.08	0.12	0.78	1.00	1.20	0.02	0.08	0.17	-0.38	0.03
## compoundVERBsdist.m	-0.14	-0.04	-0.15	-0.08	-0.01	-0.09	-0.04	0.02	-0.32	-0.03
## VERBfrac.v	0.23	0.37	-0.13	-0.04	0.07	-0.35	-0.21	-0.11	-0.42	-0.06
## NEGcount.v	-0.03	0.05	0.57	0.75	1.02	-0.03	0.02	0.09	-0.62	-0.11
## compoundVERBsdist.v	-0.20	-0.09	-0.03	0.04	0.11	-0.08	0.00	0.10	-0.18	0.06
## predsubjdist.v	-0.03	0.09	0.00	0.10	0.20	0.04	0.13	0.22	-0.17	0.17
## mamr	0.02	0.07	-0.09	0.01	0.07	-0.05	0.02	0.07	-0.10	0.16
## obj	0.00	0.05	0.03	0.08	0.16	0.58	0.83	1.17	-0.19	0.10
## predorder.v	-0.05	0.04	-0.02	0.07	0.15	0.06	0.16	0.26	-0.08	0.17
## verbalNOUNs	-0.12	-0.02	-0.28	-0.14	-0.03	-0.32	-0.18	-0.06	-0.29	0.00
## NEGfrac.m	0.60	0.69	0.18	0.29	0.39	-0.31	-0.21	-0.13	-0.21	0.09
##	upper	low	PA7	upper						
## sentlen.m	0.78	-0.07	-0.02	0.05						
## sentcount	0.15	-0.09	-0.01	0.03						
## atl	0.47	0.12	0.30	0.46						
## activity	0.03	-0.19	-0.09	-0.02						
## VERBfrac.m	0.14	-0.13	-0.05	0.02						
## wordcount	0.01	-0.03	0.01	0.06						
## entropy	0.17	0.16	0.39	0.75						
## sentlen.v	0.34	-0.13	-0.02	0.08						
## predsubjdist.m	1.44	-0.37	-0.04	0.20						
## compoundVERBs	0.18	-0.12	-0.04	0.05						
## passives	0.08	-0.15	-0.09	0.00						
## predobjdist.m	1.09	-0.20	0.00	0.14						
## literary	0.10	-0.06	0.06	0.23						
## verbdist	0.70	-0.13	-0.04	0.01						
## maentropy	0.42	0.42	0.82	1.37						
## predorder.m	1.19	-0.15	0.07	0.15						
## hapaxes	0.16	0.15	0.29	0.47						
## VERBcomp	0.13	-0.05	0.04	0.15						
## NOUNcount.v	0.11	-0.18	-0.03	0.22						
## subj	0.28	-0.42	-0.14	0.02						
## NOUNcount.m	0.60	-0.03	0.07	0.18						
## predobjdist.v	0.47	-0.10	0.02	0.13						
## NEGcount.m	0.32	0.02	0.09	0.17						
## compoundVERBsdist.m	0.43	-0.26	-0.14	-0.06						
## VERBfrac.v	0.31	-0.12	0.06	0.31						
## NEGcount.v	0.24	-0.01	0.07	0.20						
## compoundVERBsdist.v	0.41	-0.18	-0.03	0.09						
## predsubjdist.v	0.72	-0.11	0.03	0.13						
## mamr	0.28	-0.49	-0.17	-0.02						
## obj	0.54	-0.08	-0.02	0.04						
## predorder.v	0.51	-0.05	0.08	0.19						
## verbalNOUNs	0.20	-0.09	0.04	0.16						
## NEGfrac.m	0.32	-0.32	-0.09	0.06						
##										
##	Interfactor correlations and bootstrapped confidence intervals									
##	lower	estimate	upper							
## PA1-PA2	-0.2565	0.1108	0.385							
## PA1-PA3	-0.9574	-0.5622	-0.013							
## PA1-PA5	-0.8353	0.3830	0.325							

```
## PA1-PA6 -0.7336 -0.3665 0.065
## PA1-PA4 -0.6260 -0.1818 0.128
## PA1-PA8 -0.5558 -0.3611 0.124
## PA1-PA7 -0.4277 -0.1660 0.118
## PA2-PA3 -0.0075 0.1702 0.315
## PA2-PA5 -0.2153 -0.2586 0.520
## PA2-PA6 -0.1282 0.2683 0.481
## PA2-PA4 -0.0718 0.2463 0.483
## PA2-PA8 -0.1865 0.0064 0.447
## PA2-PA7 -0.1392 0.1785 0.330
## PA3-PA5 -0.3360 -0.3255 0.741
## PA3-PA6 -0.1740 0.3000 0.700
## PA3-PA4 -0.1590 0.3241 0.664
## PA3-PA8 -0.0703 0.2427 0.540
## PA3-PA7 -0.2005 0.1085 0.398
## PA5-PA6 -0.3612 -0.3378 0.713
## PA5-PA4 -0.3328 -0.2304 0.617
## PA5-PA8 -0.2715 -0.3838 0.476
## PA5-PA7 -0.2789 -0.1659 0.327
## PA6-PA4 -0.1805 0.3221 0.512
## PA6-PA8 -0.1871 0.1114 0.425
## PA6-PA7 -0.2059 0.0710 0.322
## PA4-PA8 -0.2099 -0.0029 0.433
## PA4-PA7 -0.2353 0.0752 0.295
## PA8-PA7 -0.2602 -0.1047 0.296
```

Healthiness diagnostics

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

```
## # A tibble: 33 x 2
##   feats          maxload
##   <chr>          <dbl>
## 1 verbalNOUNs    0.232
## 2 compoundVERBsdist.v 0.281
## 3 literary      0.343
## 4 predsubjdist.v 0.377
## 5 NOUNcount.v    0.431
## 6 predobjdist.v  0.509
## 7 predorder.m    0.515
## 8 predorder.v    0.519
## 9 VERBfrac.v     0.549
## 10 predsubjdist.m 0.551
## # i 23 more rows
```

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

```
##      verbalNOUNs      literary compoundVERBsdist.v      VERBfrac.v
##      0.1379713      0.2423431      0.3280540      0.3524992
##      predobjdist.v      NEGfrac.m      NOUNcount.v      predobjdist.m
##      0.3939045      0.3975483      0.4064061      0.4169727
##      compoundVERBsdist.m      predsubjdist.m      sentlen.v      predsubjdist.v
##      0.4336188      0.4453023      0.4615505      0.4669617
##      predorder.v      atl      passives      subj
##      0.5370148      0.5694476      0.5733804      0.5775257
##      NEGcount.v      VERBcomp      obj      compoundVERBs
##      0.5854885      0.5958715      0.6784960      0.7020210
##      predorder.m      hapaxes      maentropy      mamr
##      0.7030408      0.7184036      0.7553256      0.7664031
##      NOUNcount.m      verbdist      entropy      wordcount
##      0.7910351      0.8118113      0.8591141      0.8864995
##      activity      VERBfrac.m      sentcount      sentlen.m
##      0.8937370      0.8998234      0.9344065      0.9365817
##      NEGcount.m
##      0.9365996
```

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

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

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

```
##      wordcount      NEGcount.m      obj      mamr
##      1.058480      1.059835      1.079227      1.183128
##      NOUNcount.m      maentropy      NEGcount.v      compoundVERBsdist.m
##      1.203656      1.249629      1.261795      1.268893
##      predobjdist.v      hapaxes      sentcount      passives
##      1.333335      1.333578      1.346796      1.350058
##      subj      sentlen.v      atl      predorder.v
##      1.372625      1.381042      1.509559      1.551827
##      verbdist      compoundVERBs      VERBfrac.m      predobjdist.m
##      1.558892      1.579530      1.616498      1.633887
##      predsubjdist.m      entropy      NEGfrac.m      VERBfrac.v
##      1.647062      1.696694      1.871425      1.926064
##      literary      sentlen.m      VERBcomp      predsubjdist.v
##      1.976897      2.244205      2.308159      2.404788
##      predorder.m      activity      NOUNcount.v      compoundVERBsdist.v
##      2.412118      2.434222      2.574050      3.113858
##      verbalNOUNs
##      3.371824
```

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

```
## [1] "sentlen.m"      "activity"      "predorder.m"
## [4] "VERBcomp"      "NOUNcount.v"      "compoundVERBsdist.v"
## [7] "predsubjdist.v"      "verbalNOUNs"
```

Feature engineering

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

```
## [1] 1.165238e-08
```

```
KMO(data_engineered_1)
```

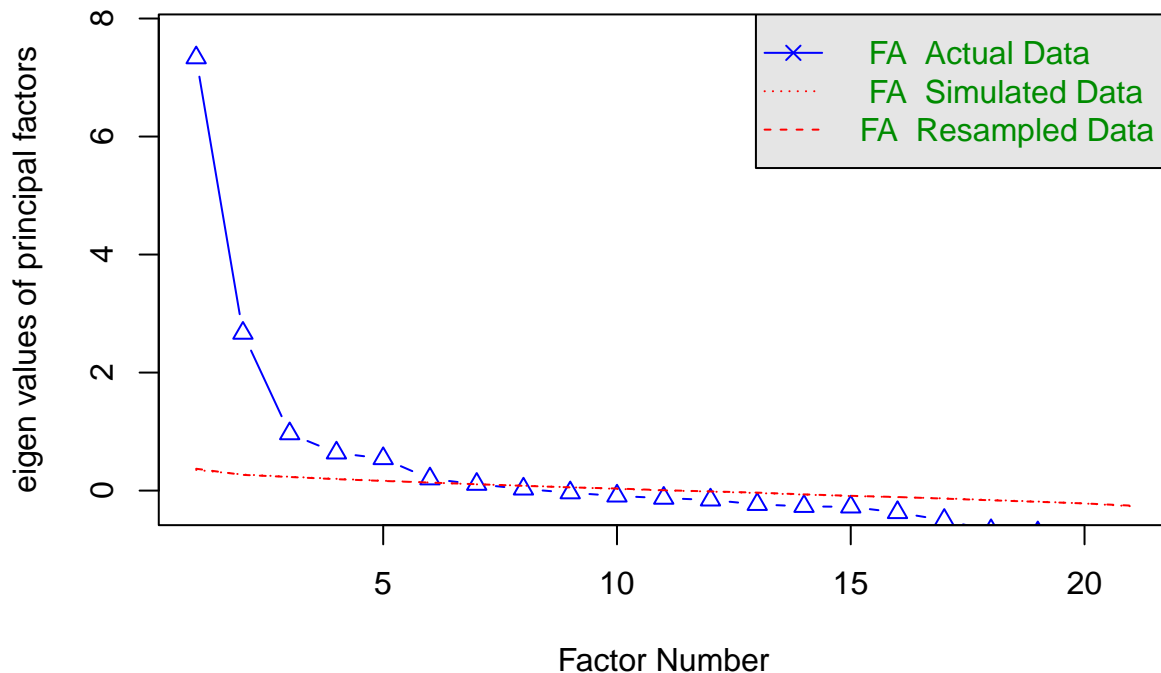
```
## Kaiser-Meyer-Olkin factor adequacy  
## Call: KMO(r = data_engineered_1)  
## Overall MSA = 0.85  
## MSA for each item =  
##      sentlen.m      sentcount          atl      activity      VERBfrac.m  
##      0.88          0.71          0.88          0.88          0.91  
##      wordcount      entropy compoundVERBs      passives      verbdist  
##      0.70          0.72          0.91          0.80          0.92  
##      maentropy      predorder.m      hapaxes      VERBcomp      subj  
##      0.60          0.88          0.80          0.88          0.95  
##      NOUNcount.m      NEGcount.m      NEGcount.v      mamr      obj  
##      0.92          0.75          0.67          0.92          0.60  
##      predorder.v  
##      0.88
```

second FA

No. of vectors

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

Parallel Analysis Scree Plots



Parallel analysis suggests that the number of factors = 6 and the number of components = NA

Model

```
set.seed(42)
```

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

```
## Factor Analysis with confidence intervals using method = fa(r = data_engineered_1, nfactors = 6, n.i
##   scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Factor Analysis using method = pa
## Call: fa(r = data_engineered_1, nfactors = 6, n.iter = 100, rotate = "promax",
##   scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##
```

	PA1	PA2	PA4	PA3	PA6	PA5	h2	u2	com
## sentlen.m	-0.73	-0.01	-0.02	0.43	0.22	0.02	0.96	0.039	1.8
## sentcount	0.19	0.92	-0.07	-0.18	-0.20	-0.06	0.92	0.084	1.3
## atl	0.67	0.03	-0.09	-0.11	0.01	0.20	0.48	0.519	1.3
## activity	0.64	-0.04	0.03	0.23	-0.39	-0.10	0.87	0.127	2.0
## VERBfrac.m	0.79	-0.04	-0.03	0.10	-0.23	-0.06	0.88	0.120	1.2
## wordcount	-0.12	0.93	0.01	0.02	0.02	0.03	0.89	0.111	1.0

```

## entropy      0.07  0.74  0.04  0.04  0.07  0.43  0.87  0.127  1.7
## compoundVERBs 0.91 -0.04 -0.02  0.02  0.38  0.02  0.62  0.375  1.3
## passives     0.10 -0.02  0.01 -0.02  0.81  0.03  0.59  0.413  1.0
## verbdist     -0.84  0.00 -0.08 -0.21  0.11 -0.09  0.78  0.218  1.2
## maentropy    -0.08 -0.05  0.01  0.04  0.05  0.87  0.78  0.215  1.0
## predorder.m  -0.72 -0.03 -0.13  0.21  0.06 -0.04  0.59  0.411  1.3
## hapaxes      0.10 -0.80 -0.03 -0.08 -0.05  0.28  0.70  0.295  1.3
## VERBcomp     0.53  0.03 -0.14  0.50 -0.15  0.07  0.60  0.402  2.3
## subj         0.73  0.12  0.01  0.00  0.19 -0.14  0.52  0.481  1.3
## NOUNcount.m  -0.94  0.05 -0.12 -0.11 -0.05  0.02  0.80  0.204  1.1
## NEGcount.m   -0.07 -0.06  0.85  0.11  0.00  0.01  0.80  0.196  1.1
## NEGcount.v   0.16  0.07  0.81  0.02  0.01  0.00  0.68  0.316  1.1
## mamr         0.82 -0.05 -0.09  0.04  0.16 -0.21  0.72  0.275  1.2
## obj          -0.05 -0.02  0.10  0.75 -0.03  0.01  0.62  0.385  1.0
## predorder.v  -0.45  0.10 -0.01  0.27  0.08  0.02  0.35  0.654  1.8
##
##              PA1  PA2  PA4  PA3  PA6  PA5
## SS loadings      6.81 2.93 1.50 1.32 1.28 1.20
## Proportion Var    0.32 0.14 0.07 0.06 0.06 0.06
## Cumulative Var    0.32 0.46 0.54 0.60 0.66 0.72
## Proportion Explained 0.45 0.20 0.10 0.09 0.08 0.08
## Cumulative Proportion 0.45 0.65 0.75 0.84 0.92 1.00
##
## With factor correlations of
##      PA1  PA2  PA4  PA3  PA6  PA5
## PA1  1.00 0.09 -0.24 -0.09 -0.50 -0.21
## PA2  0.09 1.00  0.31  0.22  0.06  0.13
## PA4 -0.24 0.31  1.00  0.26  0.31  0.22
## PA3 -0.09 0.22  0.26  1.00 -0.04 -0.07
## PA6 -0.50 0.06  0.31 -0.04  1.00  0.04
## PA5 -0.21 0.13  0.22 -0.07  0.04  1.00
##
## Mean item complexity = 1.4
## Test of the hypothesis that 6 factors are sufficient.
##
## df null model = 210 with the objective function = 18.27 with Chi Square = 13594.25
## df of the model are 99 and the objective function was 1.66
##
## The root mean square of the residuals (RMSR) is 0.02
## The df corrected root mean square of the residuals is 0.04
##
## The harmonic n.obs is 753 with the empirical chi square 195.42 with prob < 2.7e-08
## The total n.obs was 753 with Likelihood Chi Square = 1227.08 with prob < 7.1e-194
##
## Tucker Lewis Index of factoring reliability = 0.82
## RMSEA index = 0.123 and the 90 % confidence intervals are 0.117 0.129
## BIC = 571.3
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##              PA1  PA2  PA4  PA3  PA6  PA5
## Correlation of (regression) scores with factors 0.99 0.98 0.93 0.95 0.91 0.92
## Multiple R square of scores with factors        0.97 0.96 0.87 0.90 0.83 0.85
## Minimum correlation of possible factor scores    0.94 0.92 0.73 0.80 0.66 0.70
##

```

```

## Coefficients and bootstrapped confidence intervals
##          low  PA1 upper  low  PA2 upper  low  PA4 upper  low  PA3
## sentlen.m -0.78 -0.73 -0.66 -0.03 -0.01  0.02 -0.04 -0.02  0.04  0.36  0.43
## sentcount  0.14  0.19  0.23  0.88  0.92  0.96 -0.14 -0.07  0.00 -0.24 -0.18
## atl        0.58  0.67  0.75 -0.03  0.03  0.08 -0.23 -0.09  0.05 -0.25 -0.11
## activity   0.55  0.64  0.71 -0.07 -0.04 -0.01 -0.03  0.03  0.09  0.18  0.23
## VERBfrac.m 0.70  0.79  0.85 -0.08 -0.04 -0.01 -0.09 -0.03  0.05  0.03  0.10
## wordcount  -0.16 -0.12 -0.08  0.90  0.93  0.96 -0.02  0.01  0.05 -0.02  0.02
## entropy    0.02  0.07  0.11  0.71  0.74  0.78 -0.01  0.04  0.10  0.00  0.04
## compoundVERBs 0.79  0.91  0.98 -0.09 -0.04  0.01 -0.09 -0.02  0.07 -0.07  0.02
## passives   0.00  0.10  0.17 -0.06 -0.02  0.02 -0.06  0.01  0.12 -0.12 -0.02
## verbdist   -0.90 -0.84 -0.75 -0.03  0.00  0.02 -0.14 -0.08 -0.02 -0.29 -0.21
## maentropy  -0.13 -0.08 -0.05 -0.07 -0.05 -0.03 -0.02  0.01  0.06 -0.01  0.04
## predorder.m -0.93 -0.72 -0.55 -0.08 -0.03  0.02 -0.30 -0.13  0.07  0.04  0.21
## hapaxes    0.04  0.10  0.15 -0.83 -0.80 -0.76 -0.09 -0.03  0.02 -0.14 -0.08
## VERBcomp   0.43  0.53  0.60 -0.02  0.03  0.07 -0.18 -0.14 -0.06  0.42  0.50
## subj       0.63  0.73  0.80  0.06  0.12  0.17 -0.06  0.01  0.09 -0.06  0.00
## NOUNcount.m -1.00 -0.94 -0.83  0.01  0.05  0.08 -0.20 -0.12 -0.06 -0.17 -0.11
## NEGcount.m  -0.11 -0.07 -0.01 -0.10 -0.06 -0.03  0.76  0.85  0.94  0.07  0.11
## NEGcount.v  0.12  0.16  0.21  0.04  0.07  0.11  0.70  0.81  0.93 -0.03  0.02
## mamr       0.75  0.82  0.88 -0.10 -0.05  0.00 -0.18 -0.09  0.01 -0.05  0.04
## obj        -0.12 -0.05  0.02 -0.06 -0.02  0.02  0.03  0.10  0.19  0.67  0.75
## predorder.v -0.54 -0.45 -0.35  0.03  0.10  0.15 -0.12 -0.01  0.12  0.15  0.27
##          upper  low  PA6 upper  low  PA5 upper
## sentlen.m  0.49  0.16  0.22  0.27 -0.02  0.02  0.06
## sentcount  -0.14 -0.27 -0.20 -0.14 -0.09 -0.06 -0.02
## atl        0.01 -0.11  0.01  0.18  0.14  0.20  0.29
## activity   0.31 -0.56 -0.39 -0.28 -0.14 -0.10 -0.06
## VERBfrac.m 0.18 -0.34 -0.23 -0.14 -0.11 -0.06 -0.02
## wordcount  0.06 -0.03  0.02  0.06  0.00  0.03  0.06
## entropy    0.09  0.02  0.07  0.12  0.38  0.43  0.49
## compoundVERBs 0.09  0.26  0.38  0.49 -0.05  0.02  0.07
## passives   0.03  0.72  0.81  0.90 -0.03  0.03  0.07
## verbdist   -0.15 -0.01  0.11  0.27 -0.14 -0.09 -0.04
## maentropy  0.06 -0.01  0.05  0.09  0.80  0.87  0.93
## predorder.m 0.39 -0.12  0.06  0.25 -0.11 -0.04  0.03
## hapaxes    -0.04 -0.12 -0.05  0.01  0.24  0.28  0.32
## VERBcomp   0.60 -0.27 -0.15 -0.07  0.01  0.07  0.12
## subj       0.05  0.08  0.19  0.32 -0.23 -0.14 -0.07
## NOUNcount.m -0.05 -0.11 -0.05  0.03 -0.02  0.02  0.08
## NEGcount.m  0.20 -0.05  0.00  0.11 -0.02  0.01  0.06
## NEGcount.v  0.11 -0.08  0.01  0.15 -0.04  0.00  0.06
## mamr       0.11  0.03  0.16  0.31 -0.28 -0.21 -0.13
## obj        0.88 -0.14 -0.03  0.06 -0.05  0.01  0.07
## predorder.v 0.41 -0.07  0.08  0.23 -0.05  0.02  0.09
##
## Interfactor correlations and bootstrapped confidence intervals
##          lower estimate upper
## PA1-PA2 -0.0702  0.094  0.257
## PA1-PA4 -0.6226 -0.237  0.015
## PA1-PA3 -0.6426 -0.088  0.078
## PA1-PA6 -0.6048 -0.496  0.136
## PA1-PA5 -0.4244 -0.213 -0.041
## PA2-PA4 -0.0141  0.306  0.475

```



```
## PA2-PA3 -0.0437    0.221  0.381
## PA2-PA6 -0.0487    0.056  0.366
## PA2-PA5  0.0019    0.132  0.227
## PA4-PA3  0.0947    0.262  0.428
## PA4-PA6 -0.0411    0.307  0.441
## PA4-PA5  0.0067    0.223  0.353
## PA3-PA6 -0.2339   -0.036  0.339
## PA3-PA5 -0.2042   -0.072  0.266
## PA6-PA5 -0.1993    0.037  0.175
```

Healthiness diagnostics

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

```
## # A tibble: 21 x 2
##   feats      maxload
##   <chr>      <dbl>
## 1 predorder.v  0.447
## 2 VERBcomp     0.525
## 3 activity     0.643
## 4 atl         0.666
## 5 predorder.m  0.722
## 6 sentlen.m    0.726
## 7 subj        0.730
## 8 entropy     0.742
## 9 obj         0.752
## 10 VERBfrac.m  0.788
## # i 11 more rows
```

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

```
##   predorder.v      atl      subj      passives  predorder.m
##   0.3455333    0.4812761    0.5190796    0.5870862    0.5894059
##   VERBcomp      obj compoundVERBs  NEGcount.v      hapaxes
##   0.5980092    0.6150300    0.6245015    0.6842486    0.7049363
##   mamr      verbdist      maentropy  NOUNcount.m  NEGcount.m
##   0.7247364    0.7820726    0.7849233    0.7957126    0.8035536
##   entropy      activity  VERBfrac.m    wordcount    sentcount
##   0.8725579    0.8730513    0.8803732    0.8886349    0.9162827
##   sentlen.m
##   0.9605998
```

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

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

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

```
##      maentropy      wordcount      passives      obj      NEGcount.m
##      1.033820      1.037321      1.040238      1.044842      1.062761
##      NOUNcount.m      NEGcount.v      verbdist      VERBfrac.m      mamr
##      1.072662      1.100447      1.209877      1.218738      1.236964
##      predorder.m      subj      atl      sentcount      hapaxes
##      1.260679      1.269337      1.283860      1.284358      1.307339
##      compoundVERBs      entropy      predorder.v      sentlen.m      activity
##      1.335633      1.669784      1.819032      1.840314      2.017190
##      VERBcomp
##      2.339645
```

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

```
## [1] "activity" "VERBcomp"
```

Feature engineering

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

det(cor(data_engineered_2))
```

```
## [1] 5.109255e-08
```

```
KMO(data_engineered_2)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = data_engineered_2)
## Overall MSA = 0.84
## MSA for each item =
```

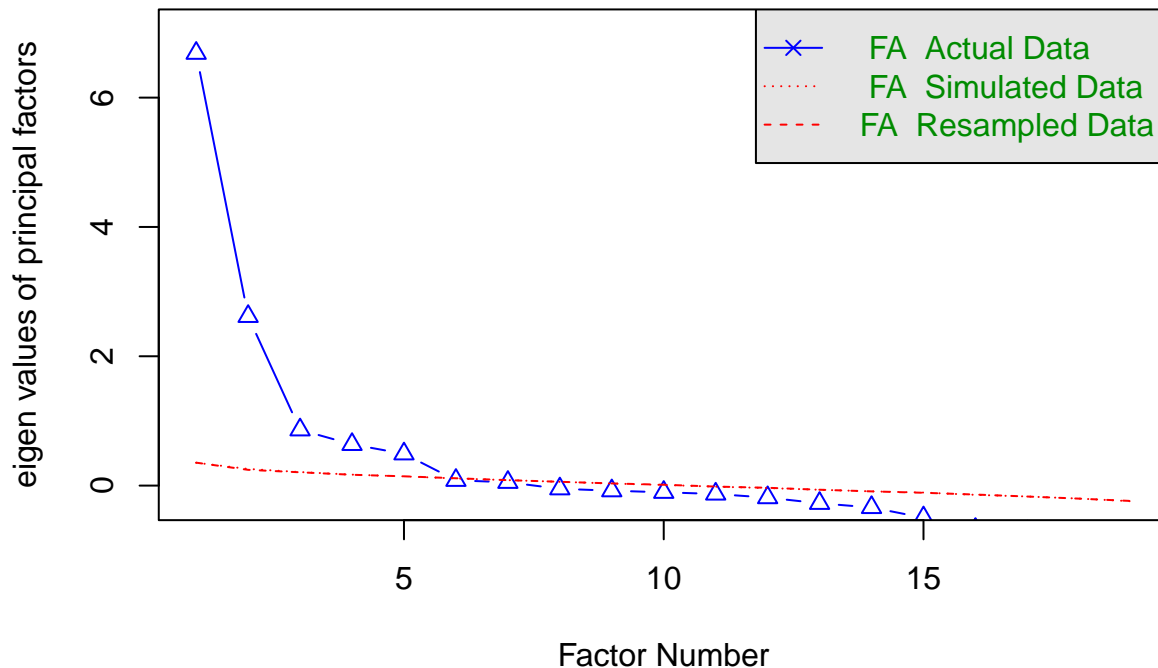
##	sentlen.m	sentcount	activity	VERBfrac.m	wordcount
##	0.84	0.71	0.90	0.90	0.70
##	entropy	compoundVERBs	passives	verbdist	maentropy
##	0.73	0.91	0.80	0.92	0.62
##	predorder.m	hapaxes	VERBcomp	subj	NOUNcount.m
##	0.89	0.79	0.88	0.94	0.92
##	NEGcount.m	NEGcount.v	mamr	obj	
##	0.73	0.67	0.90	0.57	

third FA

No. of vectors

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

Parallel Analysis Scree Plots



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

Model

```
set.seed(42)
```

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

```
## Factor Analysis with confidence intervals using method = fa(r = data_engineered_2, nfactors = 5, n.i
##   scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Factor Analysis using method = pa
## Call: fa(r = data_engineered_2, nfactors = 5, n.iter = 100, rotate = "promax",
##   scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##
```

	PA1	PA2	PA3	PA4	PA5	h2	u2	com
## sentlen.m	-0.88	0.03	0.23	0.09	-0.12	0.90	0.097	1.2
## sentcount	0.25	0.90	-0.23	0.04	0.03	0.89	0.109	1.3
## activity	0.61	-0.05	-0.04	0.52	-0.02	0.88	0.115	2.0
## VERBfrac.m	0.76	-0.05	-0.08	0.32	-0.02	0.90	0.105	1.4
## wordcount	-0.14	0.94	0.03	0.00	0.03	0.89	0.106	1.0
## entropy	0.02	0.75	0.17	0.02	0.41	0.85	0.151	1.7

```

## compoundVERBs  0.81 -0.03  0.13 -0.16 -0.10 0.56 0.437 1.2
## passives      -0.04  0.00  0.27 -0.51 -0.13 0.35 0.649 1.7
## verbdist      -0.75  0.00 -0.15 -0.33 -0.08 0.78 0.224 1.5
## maentropy     -0.16  0.00  0.19  0.03  0.68 0.54 0.461 1.3
## predorder.m   -0.78 -0.02 -0.04  0.04 -0.12 0.57 0.435 1.1
## hapaxes       0.13 -0.80 -0.03  0.00  0.31 0.73 0.268 1.4
## VERBcomp      0.30  0.05  0.03  0.58 -0.05 0.57 0.432 1.5
## subj         0.67  0.12  0.07 -0.08 -0.22 0.51 0.485 1.3
## NOUNcount.m   -0.90  0.06 -0.20 -0.12  0.03 0.81 0.190 1.1
## NEGcount.m     0.03 -0.07  0.87 -0.05  0.16 0.70 0.295 1.1
## NEGcount.v     0.28  0.05  0.81 -0.10  0.15 0.61 0.390 1.3
## mamr          0.74 -0.05 -0.04 -0.01 -0.32 0.72 0.276 1.4
## obj          -0.30  0.01  0.40  0.59 -0.14 0.60 0.403 2.5
##
##
##          PA1  PA2  PA3  PA4  PA5
## SS loadings      5.92 2.92 1.81 1.71 1.02
## Proportion Var    0.31 0.15 0.10 0.09 0.05
## Cumulative Var    0.31 0.47 0.56 0.65 0.70
## Proportion Explained 0.44 0.22 0.14 0.13 0.08
## Cumulative Proportion 0.44 0.66 0.80 0.92 1.00
##
## With factor correlations of
##          PA1  PA2  PA3  PA4  PA5
## PA1  1.00 0.10 -0.32  0.35 -0.13
## PA2  0.10 1.00  0.33  0.10  0.03
## PA3 -0.32 0.33  1.00  0.00 -0.09
## PA4  0.35 0.10  0.00  1.00 -0.17
## PA5 -0.13 0.03 -0.09 -0.17  1.00
##
## Mean item complexity = 1.4
## Test of the hypothesis that 5 factors are sufficient.
##
## df null model = 171 with the objective function = 16.79 with Chi Square = 12505.47
## df of the model are 86 and the objective function was 1.95
##
## The root mean square of the residuals (RMSR) is 0.03
## The df corrected root mean square of the residuals is 0.04
##
## The harmonic n.obs is 753 with the empirical chi square 259.11 with prob < 3e-19
## The total n.obs was 753 with Likelihood Chi Square = 1443.45 with prob < 3.1e-245
##
## Tucker Lewis Index of factoring reliability = 0.78
## RMSEA index = 0.145 and the 90 % confidence intervals are 0.138 0.151
## BIC = 873.78
## Fit based upon off diagonal values = 0.99
## Measures of factor score adequacy
##
##          PA1  PA2  PA3  PA4  PA5
## Correlation of (regression) scores with factors 0.98 0.98 0.93 0.93 0.89
## Multiple R square of scores with factors 0.97 0.95 0.87 0.86 0.79
## Minimum correlation of possible factor scores 0.94 0.91 0.74 0.72 0.57
##
## Coefficients and bootstrapped confidence intervals
##          low  PA1 upper  low  PA2 upper  low  PA3 upper  low  PA4
## sentlen.m   -0.97 -0.88 -0.78 -0.01  0.03  0.06  0.18  0.23  0.31  0.03  0.09

```

## sentcount	0.21	0.25	0.31	0.86	0.90	0.94	-0.29	-0.23	-0.17	0.00	0.04
## activity	0.54	0.61	0.70	-0.08	-0.05	-0.01	-0.09	-0.04	0.01	0.42	0.52
## VERBfrac.m	0.69	0.76	0.85	-0.08	-0.05	-0.02	-0.12	-0.08	-0.03	0.25	0.32
## wordcount	-0.16	-0.14	-0.10	0.91	0.94	0.96	0.00	0.03	0.06	-0.03	0.00
## entropy	-0.03	0.02	0.07	0.70	0.75	0.79	0.12	0.17	0.24	-0.04	0.02
## compoundVERBs	0.74	0.81	0.90	-0.09	-0.03	0.02	0.05	0.13	0.22	-0.25	-0.16
## passives	-0.14	-0.04	0.06	-0.06	0.00	0.08	0.17	0.27	0.39	-0.64	-0.51
## verbdist	-0.84	-0.75	-0.70	-0.03	0.00	0.03	-0.24	-0.15	-0.07	-0.41	-0.33
## maentropy	-0.25	-0.16	-0.09	-0.05	0.00	0.04	0.12	0.19	0.29	-0.06	0.03
## predorder.m	-0.88	-0.78	-0.71	-0.07	-0.02	0.03	-0.13	-0.04	0.07	-0.05	0.04
## hapaxes	0.09	0.13	0.16	-0.83	-0.80	-0.77	-0.08	-0.03	0.02	-0.05	0.00
## VERBcomp	0.23	0.30	0.40	0.00	0.05	0.11	-0.04	0.03	0.12	0.45	0.58
## subj	0.61	0.67	0.75	0.05	0.12	0.17	0.01	0.07	0.13	-0.13	-0.08
## NOUNcount.m	-0.97	-0.90	-0.84	0.02	0.06	0.09	-0.26	-0.20	-0.14	-0.16	-0.12
## NEGcount.m	-0.06	0.03	0.10	-0.10	-0.07	-0.02	0.79	0.87	0.98	-0.12	-0.05
## NEGcount.v	0.18	0.28	0.36	0.01	0.05	0.11	0.74	0.81	0.92	-0.16	-0.10
## mamr	0.67	0.74	0.83	-0.09	-0.05	0.00	-0.11	-0.04	0.02	-0.07	-0.01
## obj	-0.40	-0.30	-0.21	-0.03	0.01	0.06	0.34	0.40	0.49	0.48	0.59
##	upper	low	PA5	upper							
## sentlen.m	0.14	-0.19	-0.12	-0.05							
## sentcount	0.09	-0.02	0.03	0.08							
## activity	0.60	-0.07	-0.02	0.03							
## VERBfrac.m	0.38	-0.08	-0.02	0.03							
## wordcount	0.04	0.00	0.03	0.07							
## entropy	0.07	0.36	0.41	0.50							
## compoundVERBs	-0.06	-0.20	-0.10	-0.01							
## passives	-0.36	-0.25	-0.13	-0.03							
## verbdist	-0.26	-0.14	-0.08	-0.02							
## maentropy	0.10	0.56	0.68	0.87							
## predorder.m	0.12	-0.26	-0.12	0.01							
## hapaxes	0.05	0.26	0.31	0.39							
## VERBcomp	0.66	-0.14	-0.05	0.02							
## subj	-0.01	-0.31	-0.22	-0.14							
## NOUNcount.m	-0.07	-0.02	0.03	0.09							
## NEGcount.m	0.04	0.08	0.16	0.26							
## NEGcount.v	0.00	0.06	0.15	0.26							
## mamr	0.05	-0.44	-0.32	-0.24							
## obj	0.65	-0.23	-0.14	-0.06							
##											
##	Interfactor correlations and bootstrapped confidence intervals										
##	lower	estimate	upper								
## PA1-PA2	0.030	0.0988	0.198								
## PA1-PA3	-0.764	-0.3161	0.430								
## PA1-PA4	-0.509	0.3484	0.631								
## PA1-PA5	-0.353	-0.1312	0.084								
## PA2-PA3	0.014	0.3284	0.467								
## PA2-PA4	-0.086	0.0957	0.399								
## PA2-PA5	-0.094	0.0314	0.188								
## PA3-PA4	-0.240	0.0017	0.150								
## PA3-PA5	-0.358	-0.0882	0.223								
## PA4-PA5	-0.347	-0.1696	0.393								

Healthiness diagnostics

```
fa_3$loadings[] %>%  
  as_tibble() %>%  
  mutate(feats = colnames(data_engineered_2)) %>%  
  select(feats, everything()) %>%  
  pivot_longer(!feats) %>%  
  mutate(value = abs(value)) %>%  
  group_by(feats) %>%  
  summarize(maxload = max(value)) %>%  
  arrange(maxload)
```

```
## # A tibble: 19 x 2  
##   feat      maxload  
##   <chr>      <dbl>  
## 1 passives    0.507  
## 2 VERBcomp    0.576  
## 3 obj         0.587  
## 4 activity    0.613  
## 5 subj        0.673  
## 6 maentropy    0.677  
## 7 mamr         0.737  
## 8 entropy     0.748  
## 9 verbdist    0.750  
## 10 VERBfrac.m 0.760  
## 11 predorder.m 0.780  
## 12 hapaxes    0.805  
## 13 NEGcount.v 0.810  
## 14 compoundVERBs 0.811  
## 15 NEGcount.m 0.865  
## 16 sentlen.m  0.877  
## 17 NOUNcount.m 0.896  
## 18 sentcount  0.900  
## 19 wordcount  0.935
```

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

```
##      passives      subj      maentropy compoundVERBs predorder.m  
## 0.3509693 0.5146531 0.5394025 0.5631851 0.5651185  
##      VERBcomp      obj      NEGcount.v  NEGcount.m      mamr  
## 0.5675835 0.5970745 0.6096614 0.7045599 0.7237484  
##      hapaxes      verbdist NOUNcount.m      entropy      activity  
## 0.7318879 0.7763733 0.8096764 0.8489362 0.8848291  
##      sentcount wordcount VERBfrac.m      sentlen.m  
## 0.8907903 0.8936181 0.8952361 0.9031961
```

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

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

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

```
##      wordcount predorder.m NEGcount.m NOUNcount.m compoundVERBs  
## 1.045774 1.061340 1.086638 1.144807 1.166381  
##      sentlen.m maentropy sentcount      subj      NEGcount.v  
## 1.206245 1.294684 1.302652 1.326581 1.347438
```

```
##      hapaxes      VERBfrac.m      mamr      verbdist      VERBcomp
##      1.351258      1.382135      1.387440      1.495842      1.549356
##      entropy      passives      activity      obj
##      1.684038      1.689338      1.974918      2.469330
```

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

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

Feature engineering

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

```
det(cor(data_engineered_3))
```

```
## [1] 9.330367e-08
```

```
KMO(data_engineered_3)
```

```
## Kaiser-Meyer-Olkin factor adequacy
```

```
## Call: KMO(r = data_engineered_3)
```

```
## Overall MSA = 0.84
```

```
## MSA for each item =
```

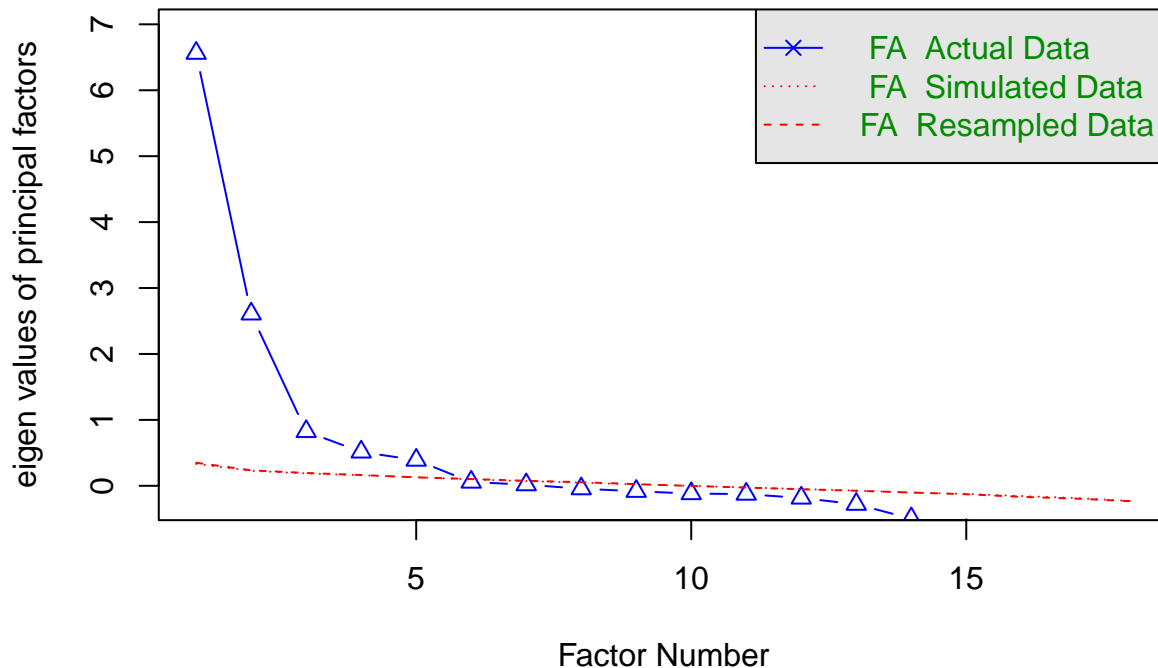
```
##      sentlen.m      sentcount      activity      VERBfrac.m      wordcount
##      0.83          0.70          0.90          0.89          0.70
##      entropy compoundVERBs      verbdist      maentropy      predorder.m
##      0.72          0.92          0.91          0.61          0.89
##      hapaxes      VERBcomp      subj      NOUNcount.m      NEGcount.m
##      0.79          0.87          0.94          0.92          0.72
##      NEGcount.v      mamr      obj
##      0.66          0.89          0.56
```

fourth FA

No. of vectors

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

Parallel Analysis Scree Plots



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

Model

```
set.seed(42)
```

```
fa_4 <- fa(
  data_engineered_3,
  nfactors = 5,
  fm = "pa",
  rotate = "promax",
  oblique.scores = TRUE,
  scores = "tenBerge",
  n.iter = 100
)
fa_4
```

```
## Factor Analysis with confidence intervals using method = fa(r = data_engineered_3, nfactors = 5, n.i
##   scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Factor Analysis using method = pa
## Call: fa(r = data_engineered_3, nfactors = 5, n.iter = 100, rotate = "promax",
##   scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##
```

	PA1	PA2	PA3	PA5	PA4	h2	u2	com
## sentlen.m	-0.86	0.04	0.10	0.37	-0.06	0.90	0.096	1.4
## sentcount	0.28	0.88	-0.15	-0.12	0.01	0.87	0.129	1.3
## activity	0.84	-0.06	-0.09	0.26	0.01	0.81	0.188	1.2
## VERBfrac.m	0.91	-0.06	-0.10	0.10	0.00	0.88	0.121	1.1
## wordcount	-0.15	0.94	0.01	0.04	0.03	0.89	0.108	1.1
## entropy	0.03	0.76	0.07	-0.01	0.43	0.87	0.133	1.6


```

## compoundVERBs  0.71 -0.02  0.10 -0.13 -0.05 0.51 0.492 1.1
## verbdist       -0.89  0.00 -0.06 -0.19 -0.11 0.78 0.222 1.1
## maentropy      -0.10 -0.02  0.04 -0.03  0.79 0.69 0.314 1.0
## predorder.m    -0.76 -0.02 -0.07  0.19 -0.10 0.56 0.437 1.2
## hapaxes        0.16 -0.81 -0.06 -0.11  0.31 0.73 0.270 1.4
## VERBcomp       0.56  0.04 -0.15  0.49  0.06 0.60 0.396 2.2
## subj          0.61  0.12  0.08 -0.06 -0.19 0.48 0.517 1.3
## NOUNcount.m    -0.93  0.05 -0.14 -0.04  0.00 0.81 0.194 1.1
## NEGcount.m     -0.05 -0.07  0.83  0.14  0.04 0.76 0.239 1.1
## NEGcount.v      0.20  0.05  0.84  0.03  0.02 0.71 0.293 1.1
## mamr           0.72 -0.04 -0.01 -0.03 -0.27 0.70 0.305 1.3
## obj           -0.07  0.01  0.14  0.78 -0.04 0.67 0.330 1.1
##
##
##          PA1  PA2  PA3  PA5  PA4
## SS loadings      6.41 2.91 1.57 1.24 1.09
## Proportion Var    0.36 0.16 0.09 0.07 0.06
## Cumulative Var    0.36 0.52 0.60 0.67 0.73
## Proportion Explained 0.48 0.22 0.12 0.09 0.08
## Cumulative Proportion 0.48 0.70 0.82 0.92 1.00
##
## With factor correlations of
##          PA1  PA2  PA3  PA5  PA4
## PA1  1.00 0.13 -0.26 0.01 -0.26
## PA2  0.13 1.00  0.30 0.15  0.10
## PA3 -0.26 0.30  1.00 0.16  0.22
## PA5  0.01 0.15  0.16 1.00  0.02
## PA4 -0.26 0.10  0.22 0.02  1.00
##
## Mean item complexity = 1.3
## Test of the hypothesis that 5 factors are sufficient.
##
## df null model = 153 with the objective function = 16.19 with Chi Square = 12062.32
## df of the model are 73 and the objective function was 1.62
##
## The root mean square of the residuals (RMSR) is 0.02
## The df corrected root mean square of the residuals is 0.04
##
## The harmonic n.obs is 753 with the empirical chi square 137.03 with prob < 8.5e-06
## The total n.obs was 753 with Likelihood Chi Square = 1205.36 with prob < 1.5e-204
##
## Tucker Lewis Index of factoring reliability = 0.8
## RMSEA index = 0.144 and the 90 % confidence intervals are 0.137 0.151
## BIC = 721.81
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##
##          PA1  PA2  PA3  PA5  PA4
## Correlation of (regression) scores with factors 0.99 0.98 0.93 0.92 0.90
## Multiple R square of scores with factors        0.97 0.95 0.86 0.84 0.81
## Minimum correlation of possible factor scores    0.94 0.91 0.72 0.68 0.62
##
## Coefficients and bootstrapped confidence intervals
##          low  PA1 upper  low  PA2 upper  low  PA3 upper  low  PA5
## sentlen.m   -0.90 -0.86 -0.82  0.00  0.04  0.07  0.06  0.10  0.15  0.33  0.37
## sentcount    0.24  0.28  0.32  0.85  0.88  0.92 -0.20 -0.15 -0.10 -0.15 -0.12

```

```

## activity      0.82  0.84  0.87 -0.10 -0.06 -0.02 -0.14 -0.09 -0.05  0.22  0.26
## VERBfrac.m    0.88  0.91  0.93 -0.10 -0.06 -0.02 -0.15 -0.10 -0.05  0.06  0.10
## wordcount     -0.17 -0.15 -0.12  0.91  0.94  0.96 -0.02  0.01  0.05  0.01  0.04
## entropy       -0.01  0.03  0.07  0.72  0.76  0.79  0.03  0.07  0.11 -0.04 -0.01
## compoundVERBs  0.66  0.71  0.77 -0.08 -0.02  0.04  0.01  0.10  0.18 -0.18 -0.13
## verbdist      -0.98 -0.89 -0.82 -0.03  0.00  0.03 -0.12 -0.06  0.00 -0.23 -0.19
## maentropy     -0.14 -0.10 -0.07 -0.04 -0.02  0.01  0.00  0.04  0.09 -0.07 -0.03
## predorder.m   -0.87 -0.76 -0.68 -0.07 -0.02  0.03 -0.16 -0.07  0.03  0.09  0.19
## hapaxes       0.12  0.16  0.20 -0.84 -0.81 -0.77 -0.11 -0.06  0.00 -0.15 -0.11
## VERBcomp      0.51  0.56  0.61  0.00  0.04  0.09 -0.21 -0.15 -0.09  0.41  0.49
## subj         0.55  0.61  0.68  0.06  0.12  0.19  0.01  0.08  0.14 -0.12 -0.06
## NOUNcount.m   -0.96 -0.93 -0.90  0.01  0.05  0.09 -0.19 -0.14 -0.09 -0.09 -0.04
## NEGcount.m    -0.08 -0.05 -0.01 -0.10 -0.07 -0.03  0.76  0.83  0.90  0.10  0.14
## NEGcount.v    0.16  0.20  0.22  0.02  0.05  0.09  0.77  0.84  0.91  0.00  0.03
## mamr          0.66  0.72  0.77 -0.10 -0.04  0.01 -0.09 -0.01  0.06 -0.10 -0.03
## obj          -0.11 -0.07 -0.03 -0.02  0.01  0.05  0.10  0.14  0.20  0.71  0.78
##
##      upper    low    PA4 upper
## sentlen.m    0.43 -0.11 -0.06 -0.02
## sentcount    -0.08 -0.03  0.01  0.05
## activity     0.31 -0.04  0.01  0.05
## VERBfrac.m   0.14 -0.05  0.00  0.05
## wordcount    0.07  0.00  0.03  0.07
## entropy      0.02  0.38  0.43  0.48
## compoundVERBs -0.07 -0.12 -0.05  0.02
## verbdist     -0.14 -0.16 -0.11 -0.07
## maentropy    0.01  0.71  0.79  0.88
## predorder.m  0.29 -0.16 -0.10 -0.04
## hapaxes     -0.07  0.27  0.31  0.36
## VERBcomp     0.57  0.00  0.06  0.12
## subj         0.00 -0.29 -0.19 -0.08
## NOUNcount.m  0.00 -0.04  0.00  0.04
## NEGcount.m   0.19  0.01  0.04  0.08
## NEGcount.v   0.07 -0.03  0.02  0.06
## mamr         0.04 -0.36 -0.27 -0.20
## obj         0.85 -0.09 -0.04  0.01
##
## Interfactor correlations and bootstrapped confidence intervals
##      lower estimate upper
## PA1-PA2  0.076  0.1305  0.20
## PA1-PA3 -0.326 -0.2550 -0.14
## PA1-PA5 -0.198  0.0081  0.16
## PA1-PA4 -0.395 -0.2633 -0.13
## PA2-PA3  0.234  0.3041  0.37
## PA2-PA5  0.055  0.1474  0.23
## PA2-PA4 -0.033  0.1017  0.22
## PA3-PA5  0.021  0.1603  0.30
## PA3-PA4  0.082  0.2182  0.34
## PA5-PA4 -0.109  0.0248  0.16

```

Healthiness diagnostics

```

fa_4$loadings[] %>%
  as_tibble() %>%
  mutate(feats = colnames(data_engineered_3)) %>%

```

```

select(feats, everything()) %>%
pivot_longer(!feat) %>%
mutate(value = abs(value)) %>%
group_by(feats) %>%
summarize(maxload = max(value)) %>%
arrange(maxload)

```

```

## # A tibble: 18 x 2
##   feat      maxload
##   <chr>      <dbl>
## 1 VERBcomp    0.564
## 2 subj        0.615
## 3 compoundVERBs 0.712
## 4 mamr        0.716
## 5 entropy     0.755
## 6 predorder.m 0.761
## 7 obj         0.778
## 8 maentropy    0.788
## 9 hapaxes     0.808
## 10 NEGcount.m 0.834
## 11 NEGcount.v 0.839
## 12 activity    0.844
## 13 sentlen.m   0.856
## 14 sentcount   0.881
## 15 verbdist    0.892
## 16 VERBfrac.m 0.909
## 17 NOUNcount.m 0.929
## 18 wordcount   0.935

```

```
fa_4$communality %>% sort()
```

```

##      subj compoundVERBs predorder.m VERBcomp      obj
## 0.4833988 0.5080629 0.5628197 0.6043396 0.6698230
## maentropy      mamr NEGcount.v      hapaxes NEGcount.m
## 0.6860621 0.6950548 0.7071884 0.7297487 0.7608026
## verbdist NOUNcount.m      activity      entropy      sentcount
## 0.7779659 0.8063441 0.8124988 0.8669604 0.8710431
## VERBfrac.m wordcount      sentlen.m
## 0.8792764 0.8922747 0.9039542

```

```
fa_4$communality[fa_4$communality < 0.5] %>% names()
```

```
## [1] "subj"
```

```
fa_4$complexity %>% sort()
```

```

##      maentropy      wordcount VERBfrac.m NOUNcount.m NEGcount.m
## 1.040417 1.056721 1.057438 1.057838 1.080802
##      obj compoundVERBs NEGcount.v      verbdist predorder.m
## 1.089701 1.110837 1.120287 1.126815 1.171046
## activity      mamr      sentcount      subj      sentlen.m
## 1.227151 1.297926 1.299314 1.324019 1.415335
##      hapaxes      entropy VERBcomp
## 1.427792 1.609326 2.170189

```

```
fa_4$complexity[fa_4$complexity > 2] %>% names()
```

```
## [1] "VERBcomp"
```

Feature engineering

```
data_engineered_4 <- data_engineered_3 %>%  
  # remove low-communality features  
  select(!c(  
    subj  
  ))
```

```
det(cor(data_engineered_4))
```

```
## [1] 1.925217e-07
```

```
KMO(data_engineered_4)
```

```
## Kaiser-Meyer-Olkin factor adequacy
```

```
## Call: KMO(r = data_engineered_4)
```

```
## Overall MSA = 0.82
```

```
## MSA for each item =
```

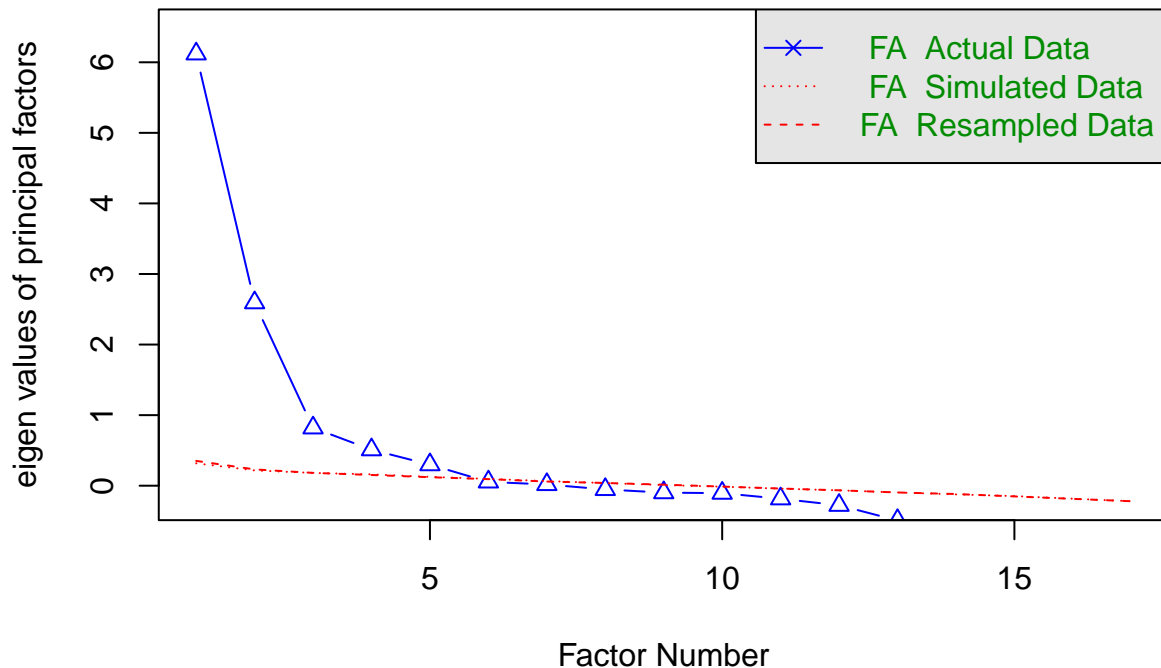
##	sentlen.m	sentcount	activity	VERBfrac.m	wordcount
##	0.82	0.69	0.89	0.88	0.70
##	entropy	compoundVERBs	verbdist	maentropy	predorder.m
##	0.72	0.91	0.91	0.59	0.88
##	hapaxes	VERBcomp	NOUNcount.m	NEGcount.m	NEGcount.v
##	0.79	0.86	0.91	0.72	0.66
##	mamr	obj			
##	0.88	0.57			

fifth FA

No. of vectors

```
fa.parallel(data_engineered_4, fm = "pa", fa = "fa", n.iter = 20)
```

Parallel Analysis Scree Plots



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

Model

```
set.seed(42)
```

```
fa_5 <- fa(
  data_engineered_4,
  nfactors = 5,
  fm = "pa",
  rotate = "promax",
  oblique.scores = TRUE,
  scores = "tenBerge",
  n.iter = 100
)
fa_5
```

```
## Factor Analysis with confidence intervals using method = fa(r = data_engineered_4, nfactors = 5, n.i
##   scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Factor Analysis using method = pa
## Call: fa(r = data_engineered_4, 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
## sentlen.m	-0.83	0.02	0.08	0.39	-0.03	0.90	0.096	1.4
## sentcount	0.27	0.89	-0.13	-0.13	-0.02	0.88	0.118	1.3
## activity	0.86	-0.04	-0.08	0.24	-0.02	0.82	0.183	1.2
## VERBfrac.m	0.92	-0.04	-0.08	0.08	-0.02	0.89	0.112	1.0
## wordcount	-0.15	0.94	0.02	0.03	0.02	0.90	0.100	1.1
## entropy	0.02	0.74	0.07	-0.01	0.39	0.84	0.159	1.5

```

## compoundVERBs  0.70  0.00  0.10 -0.14 -0.05 0.50 0.501 1.1
## verbdist      -0.90 -0.01 -0.06 -0.17 -0.09 0.78 0.222 1.1
## maentropy     -0.08 -0.05  0.00  0.00  0.90 0.84 0.164 1.0
## predorder.m  -0.74 -0.03 -0.08  0.20 -0.07 0.56 0.440 1.2
## hapaxes       0.15 -0.80 -0.06 -0.10  0.30 0.70 0.298 1.4
## VERBcomp      0.59  0.05 -0.15  0.47  0.06 0.60 0.403 2.1
## NOUNcount.m  -0.92  0.03 -0.15 -0.03  0.01 0.80 0.203 1.1
## NEGcount.m   -0.06 -0.06  0.82  0.14  0.02 0.75 0.255 1.1
## NEGcount.v    0.19  0.06  0.87  0.02 -0.02 0.74 0.261 1.1
## mamr          0.71 -0.03 -0.02 -0.04 -0.24 0.66 0.344 1.2
## obj          -0.04  0.00  0.12  0.80 -0.01 0.70 0.299 1.0
##
##
##              PA1  PA2  PA5  PA3  PA4
## SS loadings      5.98 2.89 1.56 1.25 1.16
## Proportion Var    0.35 0.17 0.09 0.07 0.07
## Cumulative Var    0.35 0.52 0.61 0.69 0.76
## Proportion Explained 0.47 0.23 0.12 0.10 0.09
## Cumulative Proportion 0.47 0.69 0.81 0.91 1.00
##
## With factor correlations of
##      PA1  PA2  PA5  PA3  PA4
## PA1  1.00 0.11 -0.26 -0.03 -0.25
## PA2  0.11 1.00  0.30  0.16  0.15
## PA5 -0.26 0.30  1.00  0.19  0.27
## PA3 -0.03 0.16  0.19  1.00  0.00
## PA4 -0.25 0.15  0.27  0.00  1.00
##
## Mean item complexity = 1.2
## Test of the hypothesis that 5 factors are sufficient.
##
## df null model = 136 with the objective function = 15.46 with Chi Square = 11527.71
## df of the model are 61 and the objective function was 1.4
##
## 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 104 with prob < 5e-04
## The total n.obs was 753 with Likelihood Chi Square = 1035.68 with prob < 3.3e-177
##
## Tucker Lewis Index of factoring reliability = 0.808
## RMSEA index = 0.146 and the 90 % confidence intervals are 0.138 0.154
## BIC = 631.61
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##
##              PA1  PA2  PA5  PA3  PA4
## Correlation of (regression) scores with factors 0.99 0.98 0.93 0.92 0.93
## Multiple R square of scores with factors        0.97 0.95 0.86 0.85 0.87
## Minimum correlation of possible factor scores    0.94 0.91 0.73 0.69 0.73
##
## Coefficients and bootstrapped confidence intervals
##      low  PA1 upper  low  PA2 upper  low  PA5 upper  low  PA3
## sentlen.m -0.87 -0.83 -0.79 -0.02 0.02 0.05 0.04 0.08 0.12 0.34 0.39
## sentcount  0.24 0.27 0.31 0.86 0.89 0.94 -0.18 -0.13 -0.09 -0.17 -0.13
## activity   0.83 0.86 0.88 -0.07 -0.04 -0.01 -0.13 -0.08 -0.02 0.19 0.24

```

```

## VERBfrac.m      0.89  0.92  0.94 -0.07 -0.04 -0.01 -0.13 -0.08 -0.02  0.03  0.08
## wordcount       -0.18 -0.15 -0.12  0.91  0.94  0.96 -0.02  0.02  0.05  0.01  0.03
## entropy         -0.02  0.02  0.06  0.71  0.74  0.78  0.03  0.07  0.11 -0.04 -0.01
## compoundVERBs   0.64  0.70  0.75 -0.06  0.00  0.05  0.03  0.10  0.17 -0.20 -0.14
## verbdist        -0.99 -0.90 -0.81 -0.04 -0.01  0.01 -0.12 -0.06  0.00 -0.22 -0.17
## maentropy       -0.12 -0.08 -0.06 -0.07 -0.05 -0.03 -0.03  0.00  0.04 -0.03  0.00
## predorder.m     -0.85 -0.74 -0.66 -0.08 -0.03  0.02 -0.18 -0.08  0.04  0.09  0.20
## hapaxes         0.11  0.15  0.18 -0.83 -0.80 -0.77 -0.11 -0.06 -0.02 -0.14 -0.10
## VERBcomp        0.55  0.59  0.65  0.00  0.05  0.09 -0.20 -0.15 -0.08  0.39  0.47
## NOUNcount.m     -0.95 -0.92 -0.89  0.00  0.03  0.06 -0.19 -0.15 -0.10 -0.08 -0.03
## NEGcount.m      -0.09 -0.06 -0.02 -0.09 -0.06 -0.02  0.74  0.82  0.92  0.10  0.14
## NEGcount.v      0.15  0.19  0.22  0.03  0.06  0.10  0.79  0.87  0.93 -0.02  0.02
## mamr            0.65  0.71  0.77 -0.08 -0.03  0.02 -0.08 -0.02  0.06 -0.10 -0.04
## obj            -0.08 -0.04  0.01 -0.03  0.00  0.03  0.07  0.12  0.18  0.73  0.80
##
##      upper    low    PA4 upper
## sentlen.m      0.43 -0.07 -0.03  0.00
## sentcount      -0.10 -0.06 -0.02  0.02
## activity        0.28 -0.06 -0.02  0.02
## VERBfrac.m     0.13 -0.07 -0.02  0.01
## wordcount       0.06 -0.01  0.02  0.05
## entropy         0.02  0.33  0.39  0.46
## compoundVERBs  -0.07 -0.12 -0.05  0.01
## verbdist        -0.12 -0.14 -0.09 -0.05
## maentropy       0.03  0.81  0.90  0.98
## predorder.m    0.29 -0.14 -0.07  0.00
## hapaxes        -0.06  0.25  0.30  0.35
## VERBcomp        0.55  0.00  0.06  0.10
## NOUNcount.m    0.02 -0.03  0.01  0.06
## NEGcount.m     0.18 -0.02  0.02  0.05
## NEGcount.v     0.06 -0.05 -0.02  0.02
## mamr           0.02 -0.32 -0.24 -0.17
## obj            0.89 -0.05 -0.01  0.04
##
## Interfactor correlations and bootstrapped confidence intervals
##      lower estimate upper
## PA1-PA2  0.039  0.1101  0.174
## PA1-PA5 -0.344 -0.2553 -0.147
## PA1-PA3 -0.259 -0.0277  0.131
## PA1-PA4 -0.384 -0.2456 -0.081
## PA2-PA5  0.223  0.2995  0.391
## PA2-PA3  0.064  0.1623  0.265
## PA2-PA4  0.063  0.1483  0.241
## PA5-PA3  0.043  0.1893  0.348
## PA5-PA4  0.139  0.2651  0.381
## PA3-PA4 -0.141  0.0033  0.150

```

Healthiness diagnostics

```

fa_5$loadings[] %>%
  as_tibble() %>%
  mutate(feats = colnames(data_engineered_4)) %>%
  select(feats, everything()) %>%
  pivot_longer(!feats) %>%
  mutate(value = abs(value)) %>%

```

```
group_by(feats) %>%
  summarize(maxload = max(value)) %>%
  arrange(maxload)
```

```
## # A tibble: 17 x 2
##   feat      maxload
##   <chr>      <dbl>
## 1 VERBcomp    0.593
## 2 compoundVERBs 0.698
## 3 mamr        0.706
## 4 predorder.m 0.743
## 5 entropy     0.745
## 6 hapaxes     0.802
## 7 obj         0.804
## 8 NEGcount.m  0.820
## 9 sentlen.m   0.833
## 10 activity    0.859
## 11 NEGcount.v  0.865
## 12 sentcount   0.895
## 13 verbdist    0.895
## 14 maentropy   0.896
## 15 VERBfrac.m  0.916
## 16 NOUNcount.m 0.920
## 17 wordcount   0.938
```

```
fa_5$communality %>% sort()
```

```
## compoundVERBs  predorder.m  VERBcomp      mamr      obj
##    0.4987070    0.5604406    0.5969918    0.6556049    0.7014127
##      hapaxes  NEGcount.v  NEGcount.m    verbdist  NOUNcount.m
##    0.7019362    0.7386412    0.7453366    0.7777822    0.7965697
##      activity  maentropy    entropy    sentcount  VERBfrac.m
##    0.8166937    0.8355523    0.8410866    0.8823881    0.8884130
##      wordcount  sentlen.m
##    0.8995476    0.9036277
```

```
fa_5$communality[fa_5$communality < 0.5] %>% names()
```

```
## [1] "compoundVERBs"
```

```
fa_5$complexity %>% sort()
```

```
##      maentropy  VERBfrac.m      obj      wordcount  NOUNcount.m
##    1.025148    1.034315    1.047293    1.052468    1.054850
##      NEGcount.m  NEGcount.v    verbdist  compoundVERBs    activity
##    1.078114    1.104370    1.104622    1.129211    1.175196
##      predorder.m      mamr    sentcount    hapaxes    sentlen.m
##    1.188548    1.248753    1.285921    1.395322    1.445528
##      entropy  VERBcomp
##    1.537031    2.084382
```

```
fa_5$complexity[fa_5$complexity > 2] %>% names()
```

```
## [1] "VERBcomp"
```


Feature engineering

```
data_engineered_5 <- data_engineered_4 %>%  
  # remove low-communality features  
  select(!c(  
    compoundVERBs  
  ))
```

```
det(cor(data_engineered_5))
```

```
## [1] 4.385204e-07
```

```
KMO(data_engineered_5)
```

```
## Kaiser-Meyer-Olkin factor adequacy
```

```
## Call: KMO(r = data_engineered_5)
```

```
## Overall MSA = 0.81
```

```
## MSA for each item =
```

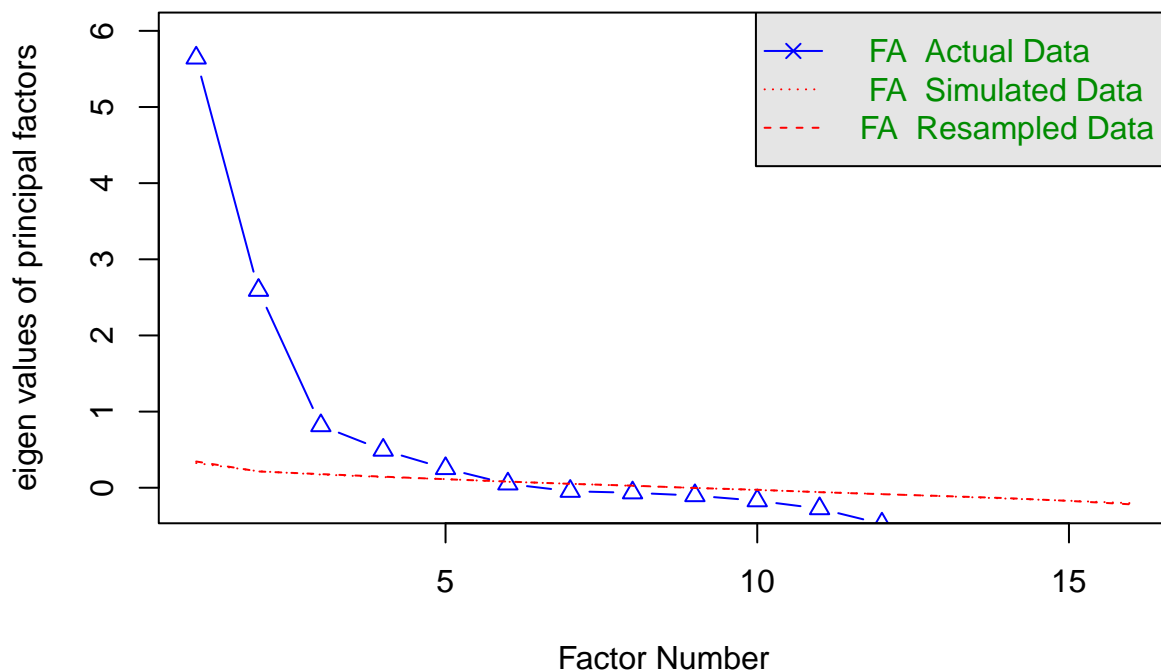
##	sentlen.m	sentcount	activity	VERBfrac.m	wordcount	entropy
##	0.81	0.69	0.88	0.87	0.70	0.73
##	verbdist	maentropy	predorder.m	hapaxes	VERBcomp	NOUNcount.m
##	0.90	0.57	0.87	0.79	0.85	0.90
##	NEGcount.m	NEGcount.v	mamr	obj		
##	0.71	0.66	0.88	0.61		

Final FA

No. of vectors

```
fa.parallel(data_engineered_5, fm = "pa", fa = "fa", n.iter = 20)
```

Parallel Analysis Scree Plots



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

Model

```
final_collist <- names(data_engineered_5)
```

```
set.seed(42)
```

```
fa_res <- fa(  
  data_engineered_5,  
  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_5, nfactors = 5, n.i  
##   scores = "tenBerge", fm = "pa", oblique.scores = TRUE)  
## Factor Analysis using method = pa  
## Call: fa(r = data_engineered_5, 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  
## sentlen.m  -0.82  0.02  0.04  0.46 -0.02 0.95 0.047 1.6  
## sentcount   0.27  0.90 -0.12 -0.16 -0.03 0.89 0.108 1.3  
## activity    0.89 -0.05 -0.06  0.20 -0.02 0.83 0.174 1.1  
## VERBfrac.m  0.92 -0.04 -0.06  0.04 -0.03 0.89 0.114 1.0  
## wordcount  -0.14  0.94  0.01  0.04  0.02 0.90 0.100 1.1  
## entropy     0.02  0.75  0.06  0.00  0.38 0.83 0.167 1.5  
## verbdist   -0.91 -0.01 -0.08 -0.13 -0.08 0.78 0.217 1.1  
## maentropy  -0.08 -0.05 -0.02  0.00  0.93 0.88 0.120 1.0  
## predorder.m -0.72 -0.03 -0.10  0.23 -0.05 0.56 0.437 1.3  
## hapaxes     0.14 -0.80 -0.06 -0.11  0.29 0.70 0.303 1.4  
## VERBcomp    0.64  0.05 -0.15  0.47  0.06 0.61 0.392 2.0  
## NOUNcount.m -0.90  0.03 -0.15  0.00  0.02 0.78 0.224 1.1  
## NEGcount.m  -0.07 -0.06  0.82  0.14  0.01 0.75 0.246 1.1  
## NEGcount.v  0.16  0.06  0.86  0.02 -0.03 0.73 0.267 1.1  
## mamr        0.69 -0.03 -0.02 -0.05 -0.24 0.63 0.369 1.3  
## obj         0.03  0.00  0.11  0.77  0.00 0.64 0.356 1.0  
##  
##  
##          PA1  PA2  PA5  PA3  PA4  
## SS loadings      5.51 2.90 1.53 1.24 1.19  
## Proportion Var    0.34 0.18 0.10 0.08 0.07  
## Cumulative Var    0.34 0.53 0.62 0.70 0.77  
## Proportion Explained 0.45 0.23 0.12 0.10 0.10  
## Cumulative Proportion 0.45 0.68 0.80 0.90 1.00  
##  
## With factor correlations of  
##  
##          PA1  PA2  PA5  PA3  PA4  
## PA1  1.00 0.11 -0.24 -0.08 -0.25  
## PA2  0.11 1.00  0.31  0.16  0.15
```

```

## PA5 -0.24 0.31 1.00 0.22 0.28
## PA3 -0.08 0.16 0.22 1.00 0.02
## PA4 -0.25 0.15 0.28 0.02 1.00
##
## Mean item complexity = 1.2
## Test of the hypothesis that 5 factors are sufficient.
##
## df null model = 120 with the objective function = 14.64 with Chi Square = 10918.9
## df of the model are 50 and the objective function was 1.12
##
## 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 69.14 with prob < 0.038
## The total n.obs was 753 with Likelihood Chi Square = 834.38 with prob < 8.6e-143
##
## Tucker Lewis Index of factoring reliability = 0.825
## RMSEA index = 0.144 and the 90 % confidence intervals are 0.136 0.153
## BIC = 503.18
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##
## Correlation of (regression) scores with factors PA1 PA2 PA5 PA3 PA4
## Multiple R square of scores with factors 0.99 0.98 0.93 0.94 0.95
## Minimum correlation of possible factor scores 0.95 0.91 0.72 0.76 0.79
##
## Coefficients and bootstrapped confidence intervals
## low PA1 upper low PA2 upper low PA5 upper low PA3
## sentlen.m -0.85 -0.82 -0.77 -0.01 0.02 0.04 0.00 0.04 0.08 0.41 0.46
## sentcount 0.24 0.27 0.31 0.86 0.90 0.93 -0.16 -0.12 -0.07 -0.20 -0.16
## activity 0.86 0.89 0.91 -0.08 -0.05 -0.02 -0.10 -0.06 0.00 0.17 0.20
## VERBfrac.m 0.90 0.92 0.95 -0.08 -0.04 -0.02 -0.11 -0.06 0.00 -0.01 0.04
## wordcount -0.17 -0.14 -0.11 0.91 0.94 0.96 -0.02 0.01 0.04 0.01 0.04
## entropy -0.03 0.02 0.05 0.72 0.75 0.78 0.02 0.06 0.11 -0.03 0.00
## verbdist -1.01 -0.91 -0.85 -0.03 -0.01 0.01 -0.13 -0.08 -0.02 -0.19 -0.13
## maentropy -0.10 -0.08 -0.05 -0.07 -0.05 -0.03 -0.05 -0.02 0.02 -0.03 0.00
## predorder.m -0.85 -0.72 -0.63 -0.09 -0.03 0.02 -0.20 -0.10 0.02 0.11 0.23
## hapaxes 0.09 0.14 0.19 -0.83 -0.80 -0.77 -0.10 -0.06 -0.01 -0.15 -0.11
## VERBcomp 0.59 0.64 0.68 0.01 0.05 0.10 -0.21 -0.15 -0.10 0.40 0.47
## NOUNcount.m -0.94 -0.90 -0.87 -0.01 0.03 0.07 -0.21 -0.15 -0.09 -0.05 0.00
## NEGcount.m -0.10 -0.07 -0.03 -0.09 -0.06 -0.02 0.73 0.82 0.91 0.10 0.14
## NEGcount.v 0.13 0.16 0.19 0.03 0.06 0.10 0.77 0.86 0.94 -0.02 0.02
## mamr 0.64 0.69 0.74 -0.07 -0.03 0.02 -0.10 -0.02 0.06 -0.13 -0.05
## obj -0.01 0.03 0.07 -0.03 0.00 0.04 0.06 0.11 0.17 0.72 0.77
## upper low PA4 upper
## sentlen.m 0.50 -0.04 -0.02 0.01
## sentcount -0.12 -0.07 -0.03 0.01
## activity 0.24 -0.06 -0.02 0.01
## VERBfrac.m 0.09 -0.07 -0.03 0.01
## wordcount 0.07 -0.01 0.02 0.05
## entropy 0.03 0.32 0.38 0.45
## verbdist -0.08 -0.13 -0.08 -0.04
## maentropy 0.04 0.85 0.93 1.02
## predorder.m 0.35 -0.10 -0.05 0.01

```

```
## hapaxes      -0.07  0.25  0.29  0.33
## VERBcomp     0.55  0.02  0.06  0.11
## NOUNcount.m  0.05 -0.02  0.02  0.06
## NEGcount.m   0.20 -0.03  0.01  0.05
## NEGcount.v   0.07 -0.06 -0.03  0.01
## mamr         0.01 -0.30 -0.24 -0.18
## obj          0.84 -0.04  0.00  0.03
##
## Interfactor correlations and bootstrapped confidence intervals
##      lower estimate upper
## PA1-PA2 -0.172    0.111  0.25
## PA1-PA5 -0.530   -0.240  0.39
## PA1-PA3 -0.360   -0.079  0.22
## PA1-PA4 -0.490   -0.249  0.33
## PA2-PA5  0.233    0.306  0.37
## PA2-PA3  0.053    0.158  0.25
## PA2-PA4  0.057    0.146  0.23
## PA5-PA3  0.077    0.225  0.37
## PA5-PA4  0.142    0.275  0.38
## PA3-PA4 -0.134    0.015  0.16
```

Healthiness diagnostics

```
fa_res$loadings[] %>%
  as_tibble() %>%
  mutate(feats = colnames(data_engineered_5)) %>%
  select(feats, everything()) %>%
  pivot_longer(!feats) %>%
  mutate(value = abs(value)) %>%
  group_by(feats) %>%
  summarize(maxload = max(value)) %>%
  arrange(maxload)
```

```
## # A tibble: 16 x 2
##   feats      maxload
##   <chr>      <dbl>
## 1 VERBcomp    0.636
## 2 mamr        0.688
## 3 predorder.m 0.724
## 4 entropy     0.747
## 5 obj         0.773
## 6 hapaxes     0.798
## 7 sentlen.m   0.817
## 8 NEGcount.m  0.819
## 9 NEGcount.v  0.860
## 10 activity   0.888
## 11 sentcount  0.897
## 12 NOUNcount.m 0.904
## 13 verbdist    0.911
## 14 VERBfrac.m  0.922
## 15 maentropy   0.927
## 16 wordcount   0.939
```

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

```
## predorder.m VERBcomp mamr obj hapaxes NEGcount.v
## 0.5631195 0.6078405 0.6310865 0.6443121 0.6967551 0.7325530
## NEGcount.m NOUNcount.m verbdist activity entropy maentropy
## 0.7541247 0.7759391 0.7830506 0.8264822 0.8332093 0.8800070
## VERBfrac.m sentcount wordcount sentlen.m
## 0.8862010 0.8917030 0.9002307 0.9530608
```

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

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

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

```
## maentropy VERBfrac.m obj wordcount NOUNcount.m verbdist
## 1.020243 1.020505 1.044929 1.051276 1.057854 1.076180
## NEGcount.v NEGcount.m activity mamr predorder.m sentcount
## 1.085446 1.085961 1.119379 1.257157 1.266564 1.288863
## hapaxes entropy sentlen.m VERBcomp
## 1.383313 1.511799 1.582206 2.013233
```

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

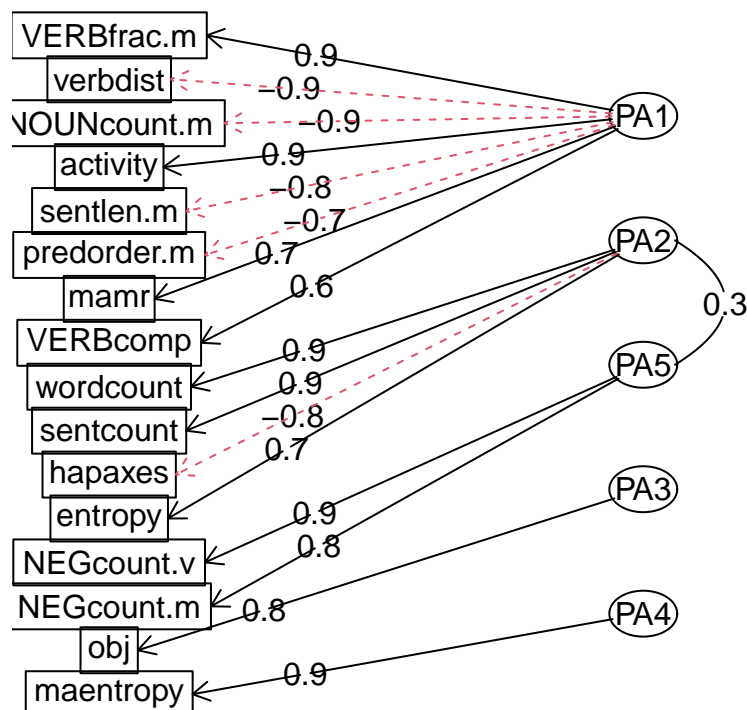
```
## [1] "VERBcomp"
```

Loadings

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

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

Factor Analysis



```
fa_res$loadings
```

```
##
## Loadings:
##          PA1    PA2    PA5    PA3    PA4
## sentlen.m -0.817          0.459
## sentcount  0.273  0.897 -0.118 -0.156
## activity   0.888          0.202
## VERBfrac.m 0.922
## wordcount  -0.143  0.939
## entropy    0.747          0.384
## verbdist   -0.911          -0.134
## maentropy           0.927
## predorder.m -0.724          -0.101  0.235
## hapaxes     0.144 -0.798          -0.108  0.290
## VERBcomp    0.636          -0.154  0.469
## NOUNcount.m -0.904          -0.149
## NEGcount.m           0.819  0.144
## NEGcount.v   0.163          0.860
## mamr         0.688          -0.240
## obj          0.111  0.773
##
##          PA1    PA2    PA5    PA3    PA4
## SS loadings  5.508 2.901 1.515 1.205 1.165
## Proportion Var 0.344 0.181 0.095 0.075 0.073
## Cumulative Var 0.344 0.526 0.620 0.696 0.768
for (i in 1:fa_res$nfactors) {
  cat("\n-----", colnames(fa_res$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(., 3))) %>%
    print()

  cat("\n")
}
```

```
##
```

```

## ----- PA1 -----
##          loading abs_l strng
## VERBfrac.m    0.922 0.922 *****
## verbdist      -0.911 0.911 *****
## NOUNcount.m   -0.904 0.904 *****
## activity       0.888 0.888 *****
## sentlen.m     -0.817 0.817 *****
## predorder.m   -0.724 0.724 *****
## mamr          0.688 0.688 ****
## VERBcomp      0.636 0.636 ****
## sentcount     0.273 0.273
## NEGcount.v    0.163 0.163
## hapaxes       0.144 0.144
## wordcount     -0.143 0.143
##
##
## ----- PA2 -----
##          loading abs_l strng
## wordcount     0.939 0.939 *****
## sentcount     0.897 0.897 *****
## hapaxes      -0.798 0.798 *****
## entropy       0.747 0.747 *****
##
##
## ----- PA5 -----
##          loading abs_l strng
## NEGcount.v    0.860 0.860 *****
## NEGcount.m    0.819 0.819 *****
## VERBcomp      -0.154 0.154
## NOUNcount.m   -0.149 0.149
## sentcount     -0.118 0.118
## obj           0.111 0.111
## predorder.m   -0.101 0.101
##
##
## ----- PA3 -----
##          loading abs_l strng
## obj           0.773 0.773 *****
## VERBcomp      0.469 0.469 **
## sentlen.m     0.459 0.459 **
## predorder.m   0.235 0.235
## activity       0.202 0.202
## sentcount     -0.156 0.156
## NEGcount.m    0.144 0.144
## verbdist      -0.134 0.134
## hapaxes       -0.108 0.108
##
##
## ----- PA4 -----
##          loading abs_l strng
## maentropy     0.927 0.927 *****
## entropy       0.384 0.384 *
## hapaxes       0.290 0.290
## mamr          -0.240 0.240

```

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)
```

##	sentlen.m	sentcount	activity	VERBfrac.m	wordcount	entropy
##	0.047	0.108	0.174	0.114	0.100	0.167
##	verbdist	maentropy	predorder.m	hapaxes	VERBcomp	NOUNcount.m
##	0.217	0.120	0.437	0.303	0.392	0.224
##	NEGcount.m	NEGcount.v	mamr	obj		
##	0.246	0.267	0.369	0.356		

Distributions over factors

```
res_data <- data_factor_bind(data_clean, fa_res)
```

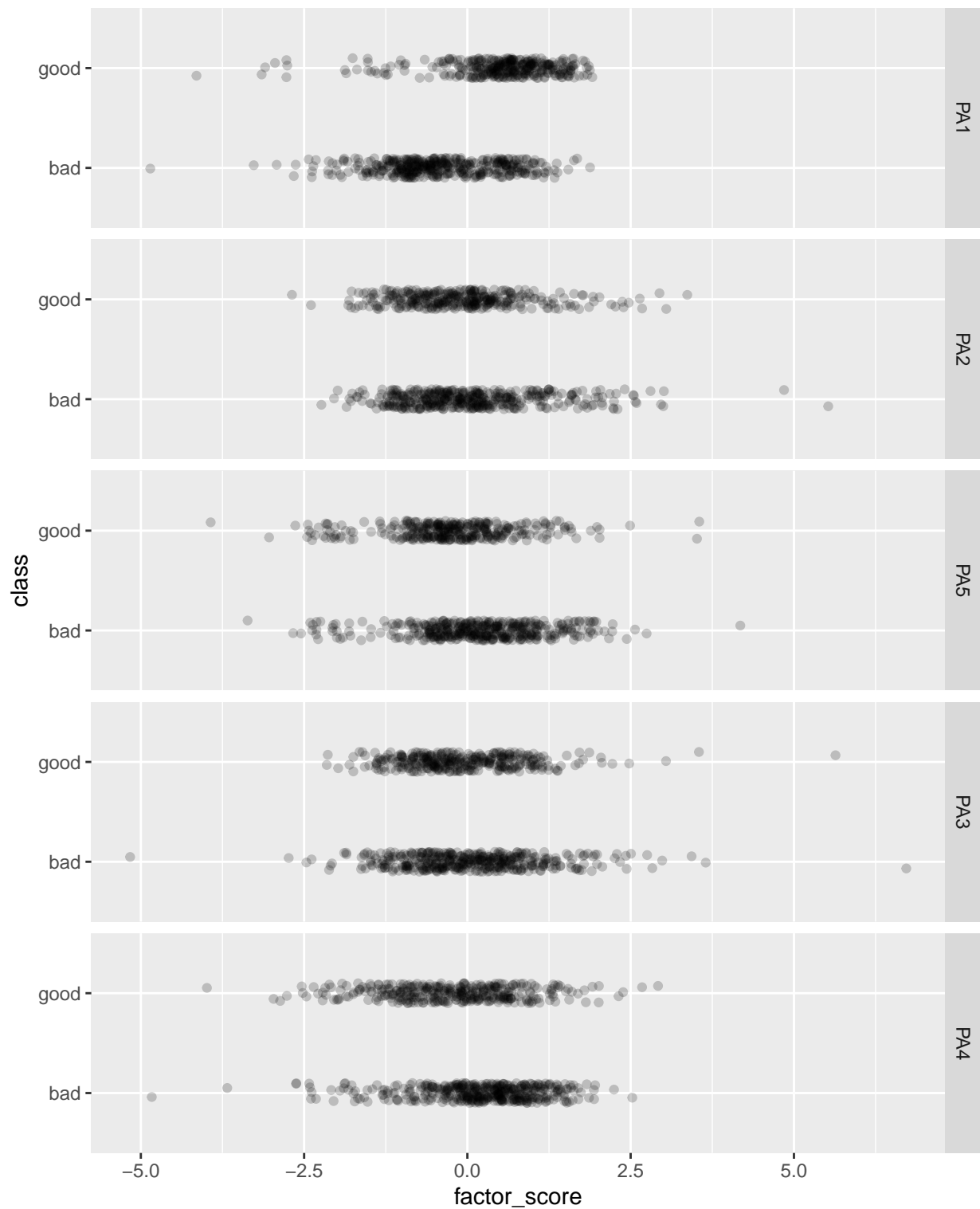
```
res_data$long %>%  
  group_by(factor) %>%  
  summarize(shapiro = shapiro.test(factor_score)$p.value)
```

```
## # A tibble: 5 x 2  
##   factor shapiro  
##   <fct>   <dbl>  
## 1 PA1    1.30e-11  
## 2 PA2    1.66e-13  
## 3 PA5    6.74e- 8
```



```
## 4 PA3      1.03e-14
## 5 PA4      1.70e- 8

res_data$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)
```

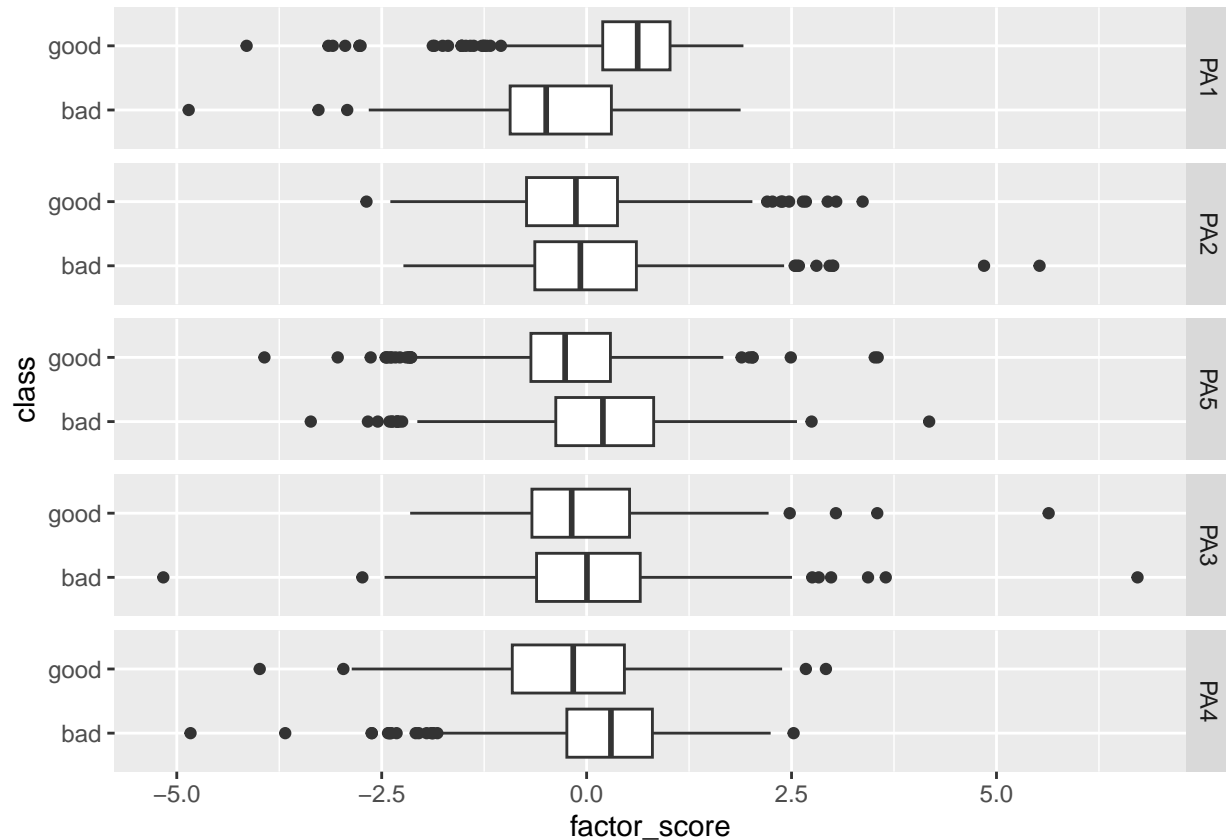


class

```
analyze_distributions(res_data$long, "class")
```

```
##
```

```
## bad good
## 2070 1695
```



```
##
## Test for the significance of differences in class over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 182.3152, df = 1, p-value = 0
##
##
##               Comparison of x by group
##               (Bonferroni)
## Col Mean-|
## Row Mean |      bad
## -----+-----
##   good | -13.50241
##       |  0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.242
##
## Test for the significance of differences in class over PA2 :
##
##   Kruskal-Wallis rank sum test
##
```

```

## data: x and group
## Kruskal-Wallis chi-squared = 2.1374, df = 1, p-value = 0.14
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |          bad
## -----+-----
##      good |    1.461991
##           |    0.1437
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.00284
##
## Test for the significance of differences in class over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 41.5404, df = 1, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |          bad
## -----+-----
##      good |    6.445187
##           |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0552
##
## Test for the significance of differences in class over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 3.4297, df = 1, p-value = 0.06
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |          bad
## -----+-----
##      good |    1.851945
##           |    0.0640
##
## alpha = 0.05
## Reject Ho if p <= alpha

```

```

## epsilon2 = 0.00456
##
## Test for the significance of differences in class over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 42.5756, df = 1, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |          bad
## -----+-----
##    good |    6.524996
##         |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0566
##
##   factor kruskal_p epsilon2
## 1    PA1 1.51e-41 0.24200
## 2    PA2 1.44e-01 0.00284
## 3    PA5 1.15e-10 0.05520
## 4    PA3 6.40e-02 0.00456
## 5    PA4 6.80e-11 0.05660
##
## p < 5e-2 found in: PA1 PA5 PA4
## p < 1e-2 found in: PA1 PA5 PA4
## p < 1e-3 found in: PA1 PA5 PA4
## p < 1e-4 found in: PA1 PA5 PA4

```

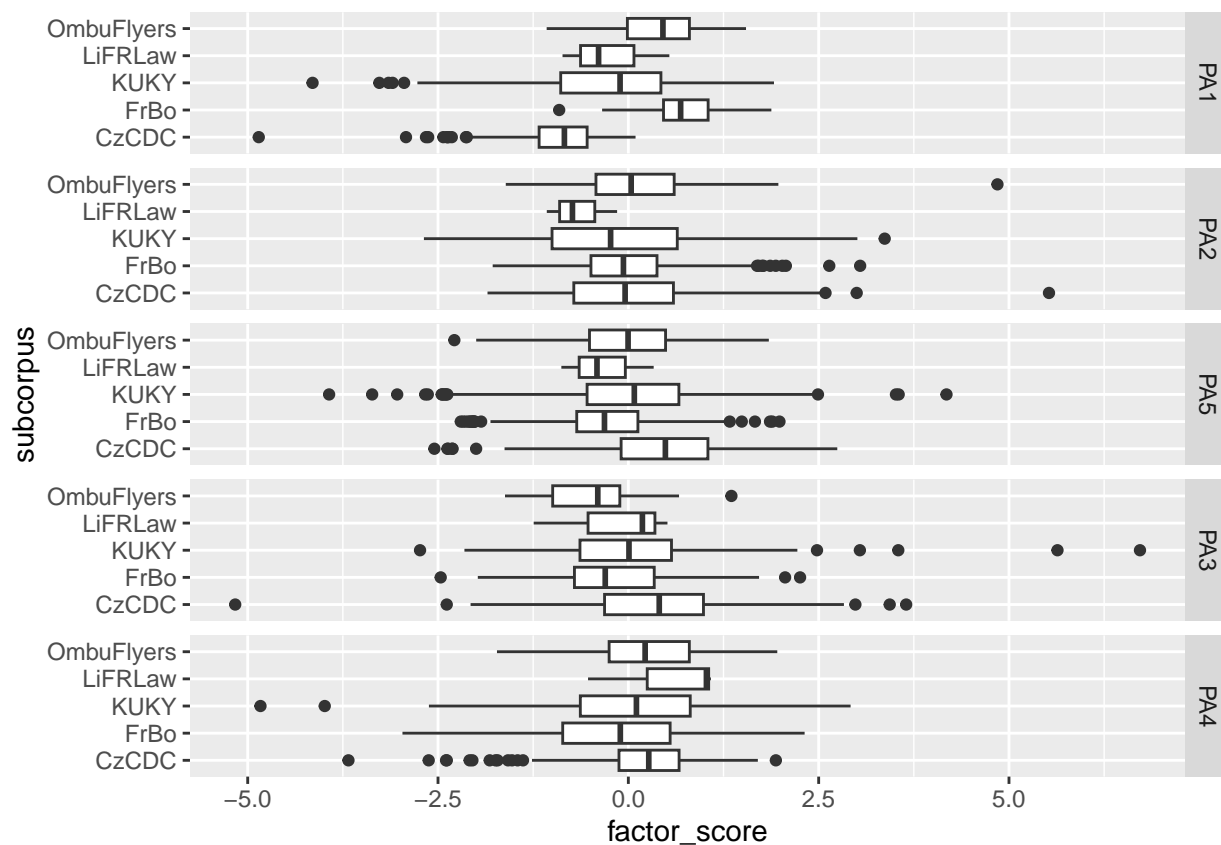
subcorpus

```
analyze_distributions(res_data$long, "subcorpus")
```

```

##
##      CzCDC      FrBo      KUKY      LiFRLaw OmbuFlyers
##      1055      1535      970          15          190

```



```
##
## Test for the significance of differences in subcorpus over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 426.8225, df = 4, p-value = 0
##
##
##           Comparison of x by group
##           (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC      FrBo      KUKY      LiFRLaw
## -----|-----
##   FrBo | -20.17725
##         |  0.0000*
##
##   KUKY | -7.614092  11.41525
##         |  0.0000*  0.0000*
##
##   LiFRLaw | -1.131913  1.975634  0.170544
##         |  1.0000  0.4820  1.0000
##
##   OmbuFlye | -7.822242  2.476388 -3.501201 -1.201138
##         |  0.0000*  0.1327  0.0046*  1.0000
##
## alpha = 0.05
```

```

## Reject Ho if p <= alpha
## epsilon2 = 0.568
##
## Test for the significance of differences in subcorpus over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 7.4477, df = 4, p-value = 0.11
##
##
##               Comparison of x by group
##               (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC      FrBo      KUKY      LiFRLaw
## -----+-----
##   FrBo | -0.327085
##         | 1.0000
##         |
##   KUKY | 1.540532  1.989644
##         | 1.0000  0.4663
##         |
## LiFRLaw | 1.366382  1.419803  1.102163
##         | 1.0000  1.0000  1.0000
##         |
## OmbuFlye | -0.967455 -0.821315 -1.824834 -1.609049
##         | 1.0000  1.0000  0.6803  1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0099
##
## Test for the significance of differences in subcorpus over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 100.0776, df = 4, p-value = 0
##
##
##               Comparison of x by group
##               (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC      FrBo      KUKY      LiFRLaw
## -----+-----
##   FrBo | 9.922486
##         | 0.0000*
##         |
##   KUKY | 4.459301 -4.838257
##         | 0.0001*  0.0000*
##         |
## LiFRLaw | 1.647206  0.121418  0.883799
##         | 0.9952  1.0000  1.0000
##         |

```

```

## OmbuFlye | 3.076617 -2.006958 0.555899 -0.692964
##          | 0.0209* 0.4475 1.0000 1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.133
##
## Test for the significance of differences in subcorpus over PA3 :
##
## Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 58.179, df = 4, p-value = 0
##
##
## Comparison of x by group
## (Bonferroni)
## Col Mean-|
## Row Mean | CzCDC FrBo KUKY LiFRLaw
## -----+-----
## FrBo | 6.783088
##      | 0.0000*
##      |
## KUKY | 3.723606 -2.575210
##      | 0.0020* 0.1002
##      |
## LiFRLaw | 0.872337 -0.171261 0.235185
##      | 1.0000 1.0000 1.0000
##      |
## OmbuFlye | 5.344900 1.949962 3.221673 0.724839
##      | 0.0000* 0.5118 0.0127* 1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0774
##
## Test for the significance of differences in subcorpus over PA4 :
##
## Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 23.0943, df = 4, p-value = 0
##
##
## Comparison of x by group
## (Bonferroni)
## Col Mean-|
## Row Mean | CzCDC FrBo KUKY LiFRLaw
## -----+-----
## FrBo | 4.331355
##      | 0.0001*
##      |
## KUKY | 1.503870 -2.592107
##      | 1.0000 0.0954

```

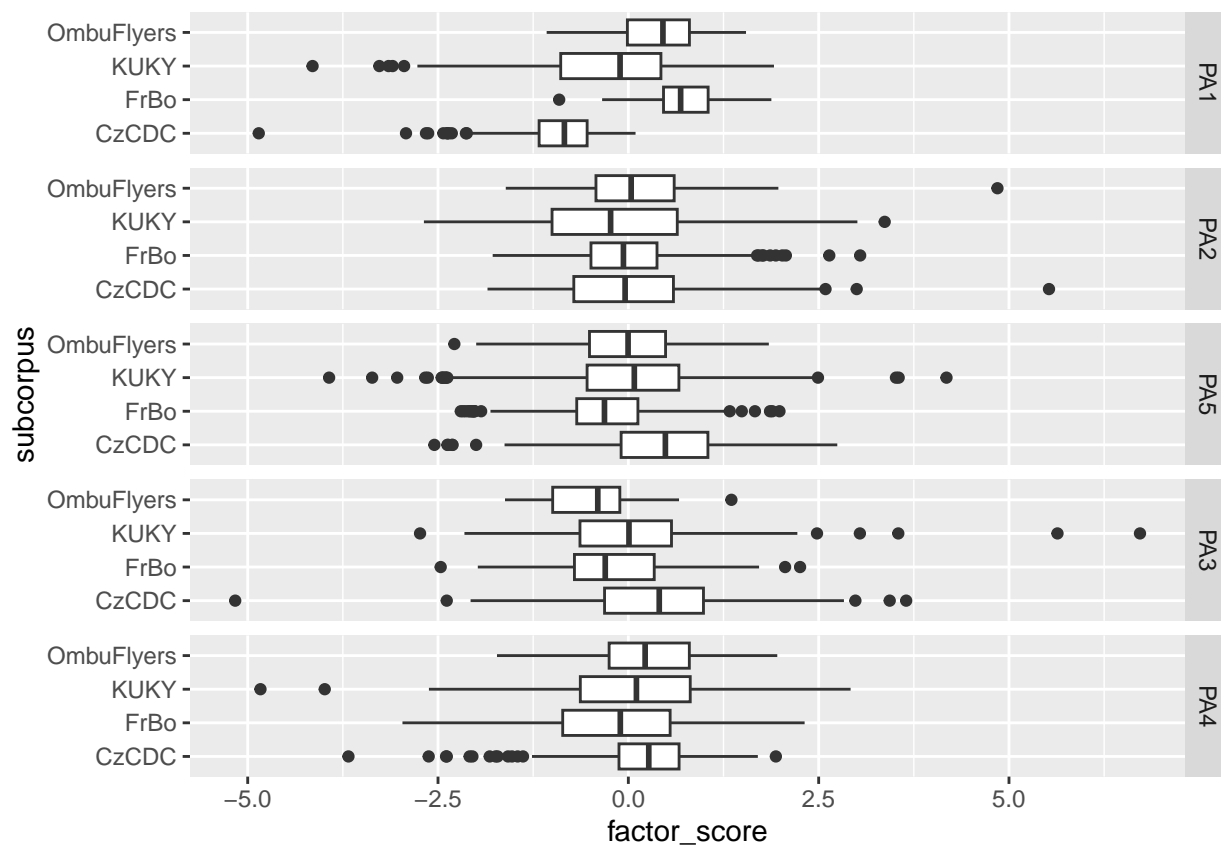


```
##
## LiFRLaw | -0.704794 -1.373961 -0.961475
##          | 1.0000 1.0000 1.0000
##
## OmbuFlye | -0.265838 -2.524736 -1.107305 0.605209
##          | 1.0000 0.1158 1.0000 1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0307
##
## factor kruskal_p epsilon2
## 1 PA1 4.45e-91 0.5680
## 2 PA2 1.14e-01 0.0099
## 3 PA5 9.47e-21 0.1330
## 4 PA3 7.00e-12 0.0774
## 5 PA4 1.21e-04 0.0307
##
## p < 5e-2 found in: PA1 PA5 PA3 PA4
## p < 1e-2 found in: PA1 PA5 PA3 PA4
## p < 1e-3 found in: PA1 PA5 PA3 PA4
## p < 1e-4 found in: PA1 PA5 PA3
```

subcorpus wo/ LiFRLaw

```
analyze_distributions(
  res_data$long %>% filter(subcorpus != "LiFRLaw"), "subcorpus"
)
```

```
##
## CzCDC FrBo KUKY LiFRLaw OmbuFlyers
## 1055 1535 970 0 190
```



```
##
## Test for the significance of differences in subcorpus over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 425.4858, df = 3, p-value = 0
##
##
##           Comparison of x by group
##           (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC      FrBo      KUKY
## -----+-----
##   FrBo | -20.15355
##         |  0.0000*
##         |
##   KUKY | -7.601464  11.40584
##         |  0.0000*  0.0000*
##         |
## OmbuFlye | -7.810550  2.476046 -3.496667
##         |  0.0000*  0.0797  0.0028*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.568
##
```

```

## Test for the significance of differences in subcorpus over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 5.5887, df = 3, p-value = 0.13
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC      FrBo      KUKY
## -----+-----
##   FrBo | -0.319407
##         | 1.0000
##         |
##   KUKY | 1.534630  1.975756
##         | 0.7492  0.2891
##         |
## OmbuFlye | -0.962888 -0.820627 -1.816987
##         | 1.0000  1.0000  0.4153
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.00746
##
## Test for the significance of differences in subcorpus over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 99.4143, df = 3, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC      FrBo      KUKY
## -----+-----
##   FrBo | 9.919546
##         | 0.0000*
##         |
##   KUKY | 4.454987 -4.840068
##         | 0.0001*  0.0000*
##         |
## OmbuFlye | 3.075538 -2.006535  0.557246
##         | 0.0126*  0.2688  1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.133
##
## Test for the significance of differences in subcorpus over PA3 :
##

```

```

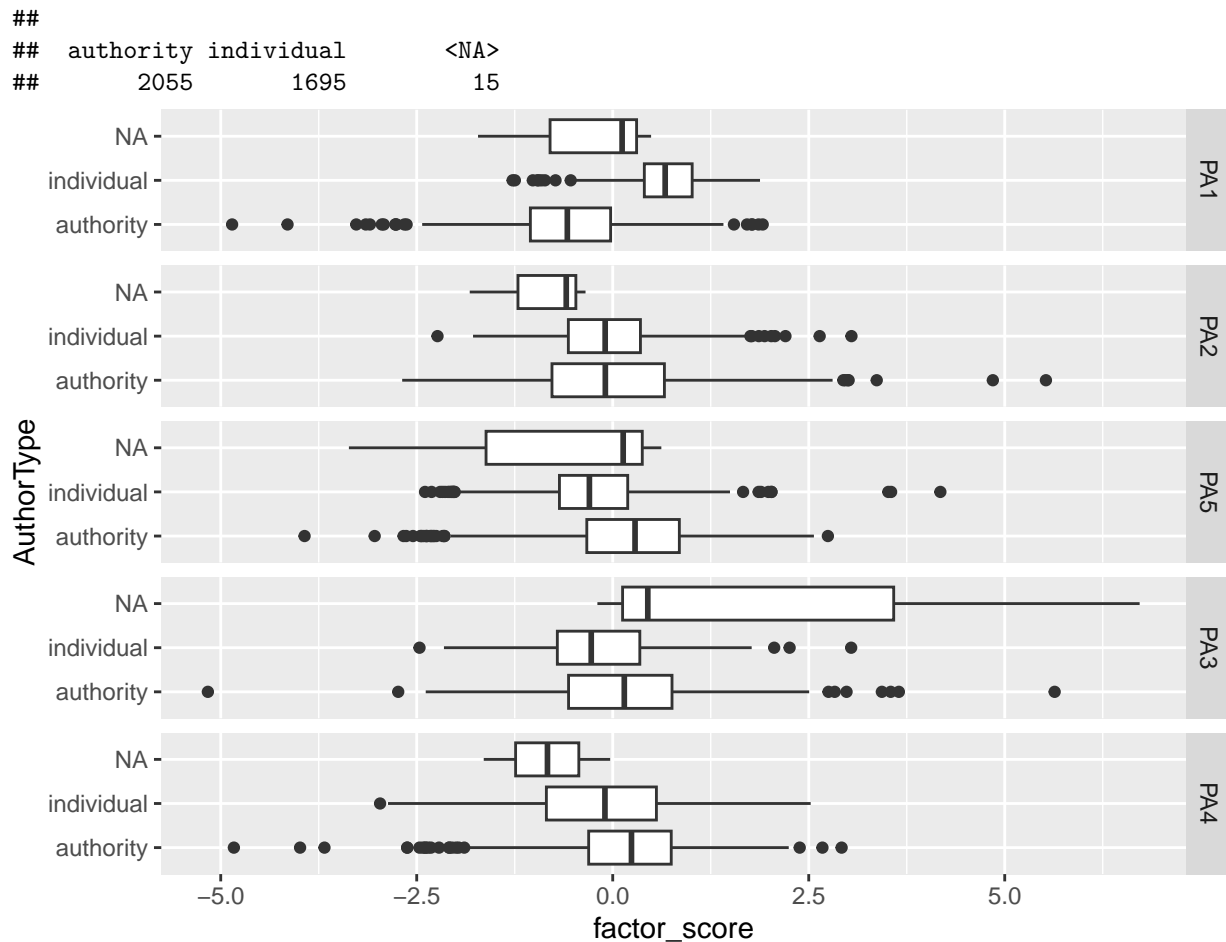
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 58.1077, df = 3, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC      FrBo      KUKY
## -----+-----
##   FrBo |      6.781070
##         |      0.0000*
##         |
##   KUKY |      3.721153  -2.575903
##         |      0.0012*      0.0600
##         |
## OmbuFlye |      5.343575   1.949653   3.221732
##         |      0.0000*      0.3073   0.0076*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0776
##
## Test for the significance of differences in subcorpus over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 22.033, df = 3, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      CzCDC      FrBo      KUKY
## -----+-----
##   FrBo |      4.336165
##         |      0.0001*
##         |
##   KUKY |      1.510151  -2.589985
##         |      0.7860      0.0576
##         |
## OmbuFlye |     -0.263124  -2.524456  -1.108132
##         |      1.0000      0.0695      1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0294
##
##   factor kruskal_p epsilon2
## 1   PA1  6.67e-92  0.56800
## 2   PA2  1.33e-01  0.00746
## 3   PA5  2.08e-21  0.13300

```

```
## 4    PA3  1.49e-12  0.07760
## 5    PA4  6.42e-05  0.02940
##
## 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(res_data$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 = 321.1551, df = 1, p-value = 0
##
##
##           Comparison of x by group
##           (Bonferroni)
## Col Mean-|
```

```

## Row Mean |   authorit
## -----+-----
## individu |  -17.92080
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.427
##
## Test for the significance of differences in AuthorType over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.0653, df = 1, p-value = 0.8
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## -----+-----
## individu |    0.255510
##          |    0.7983
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 8.68e-05
##
## Test for the significance of differences in AuthorType over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 64.9803, df = 1, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## -----+-----
## individu |    8.061033
##          |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0864
##
## Test for the significance of differences in AuthorType over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group

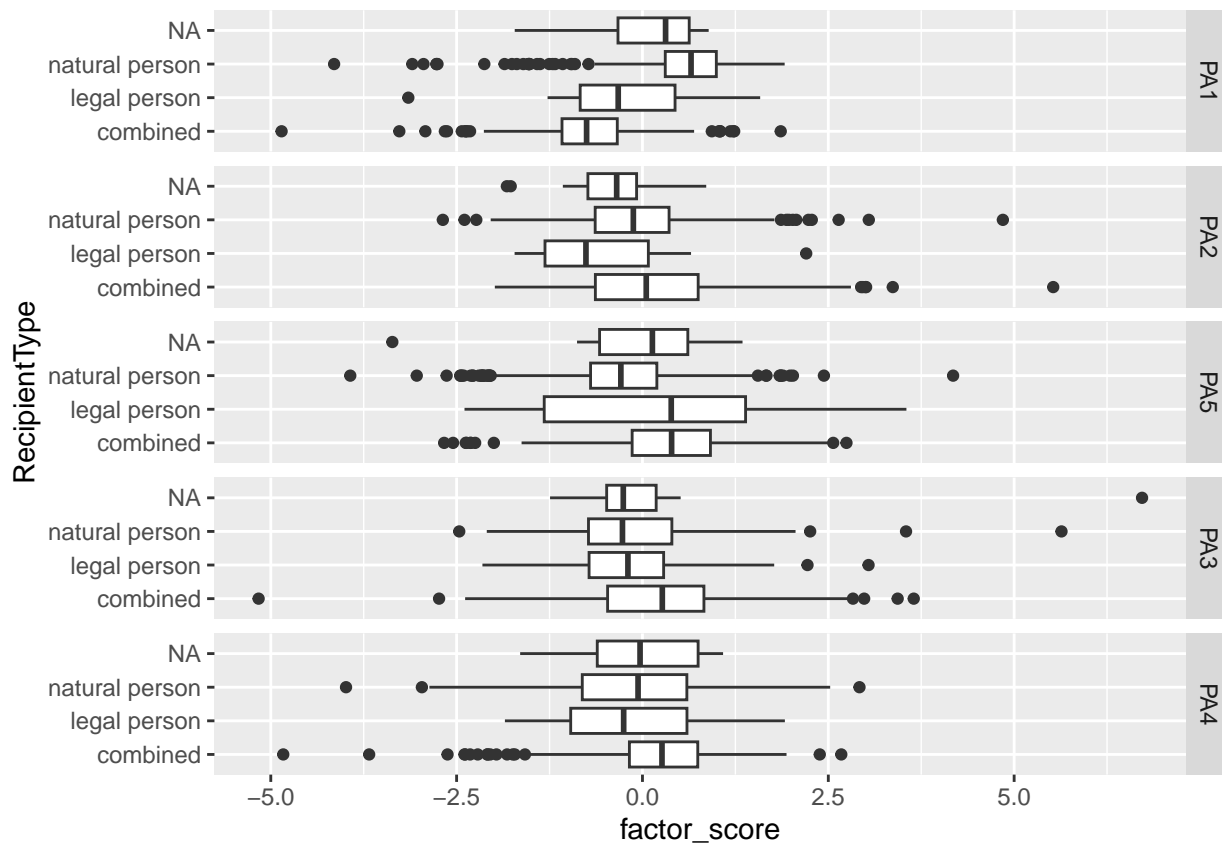
```

```
## Kruskal-Wallis chi-squared = 22.0195, df = 1, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## -----+-----
## individu |   4.692491
##          |   0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0293
##
## Test for the significance of differences in AuthorType over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 20.4805, df = 1, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |   authorit
## -----+-----
## individu |   4.525537
##          |   0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0272
##
##   factor kruskal_p epsilon2
## 1    PA1  8.12e-72 4.27e-01
## 2    PA2  7.98e-01 8.68e-05
## 3    PA5  7.57e-16 8.64e-02
## 4    PA3  2.70e-06 2.93e-02
## 5    PA4  6.02e-06 2.72e-02
##
## p < 5e-2 found in: PA1 PA5 PA3 PA4
## p < 1e-2 found in: PA1 PA5 PA3 PA4
## p < 1e-3 found in: PA1 PA5 PA3 PA4
## p < 1e-4 found in: PA1 PA5 PA3 PA4
```

RecipientType

```
analyze_distributions(res_data$long, "RecipientType")
```

```
##
##      combined   legal person natural person      <NA>
##          1520           115           2065           65
```



```
##
## Test for the significance of differences in RecipientType over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 348.3481, df = 2, p-value = 0
##
##
##           Comparison of x by group
##           (Bonferroni)
## Col Mean-|
## Row Mean |   combined   legal pe
## -----+-----
## legal pe |  -2.389864
##           |      0.0506
##           |
## natural  |  -18.61263  -4.152891
##           |      0.0000*   0.0001*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.463
##
## Test for the significance of differences in RecipientType over PA2 :
##
##   Kruskal-Wallis rank sum test
```



```

##
## data: x and group
## Kruskal-Wallis chi-squared = 16.1034, df = 2, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |    combined    legal pe
## -----+-----
## legal pe |    3.636925
##          |    0.0008*
##          |
## natural  |    2.393265  -2.826981
##          |    0.0501    0.0141*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0214
##
## Test for the significance of differences in RecipientType over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 105.1664, df = 2, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |    combined    legal pe
## -----+-----
## legal pe |    1.141705
##          |    0.7607
##          |
## natural  |   10.21253   2.449825
##          |    0.0000*    0.0429*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.14
##
## Test for the significance of differences in RecipientType over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 33.1119, df = 2, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|

```

```

## Row Mean |    combined    legal pe
## -----+-----
## legal pe |    1.704506
##          |    0.2649
##          |
## natural  |    5.726649    0.299415
##          |    0.0000*    1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.044
##
## Test for the significance of differences in RecipientType over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 19.5102, df = 2, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |    combined    legal pe
## -----+-----
## legal pe |    1.786618
##          |    0.2220
##          |
## natural  |    4.316515   -0.280867
##          |    0.0000*    1.0000
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0259
##
##   factor kruskal_p epsilon2
## 1   PA1  2.28e-76  0.4630
## 2   PA2  3.19e-04  0.0214
## 3   PA5  1.46e-23  0.1400
## 4   PA3  6.45e-08  0.0440
## 5   PA4  5.80e-05  0.0259
##
## p < 5e-2 found in: PA1 PA2 PA5 PA3 PA4
## p < 1e-2 found in: PA1 PA2 PA5 PA3 PA4
## p < 1e-3 found in: PA1 PA2 PA5 PA3 PA4
## p < 1e-4 found in: PA1 PA5 PA3 PA4

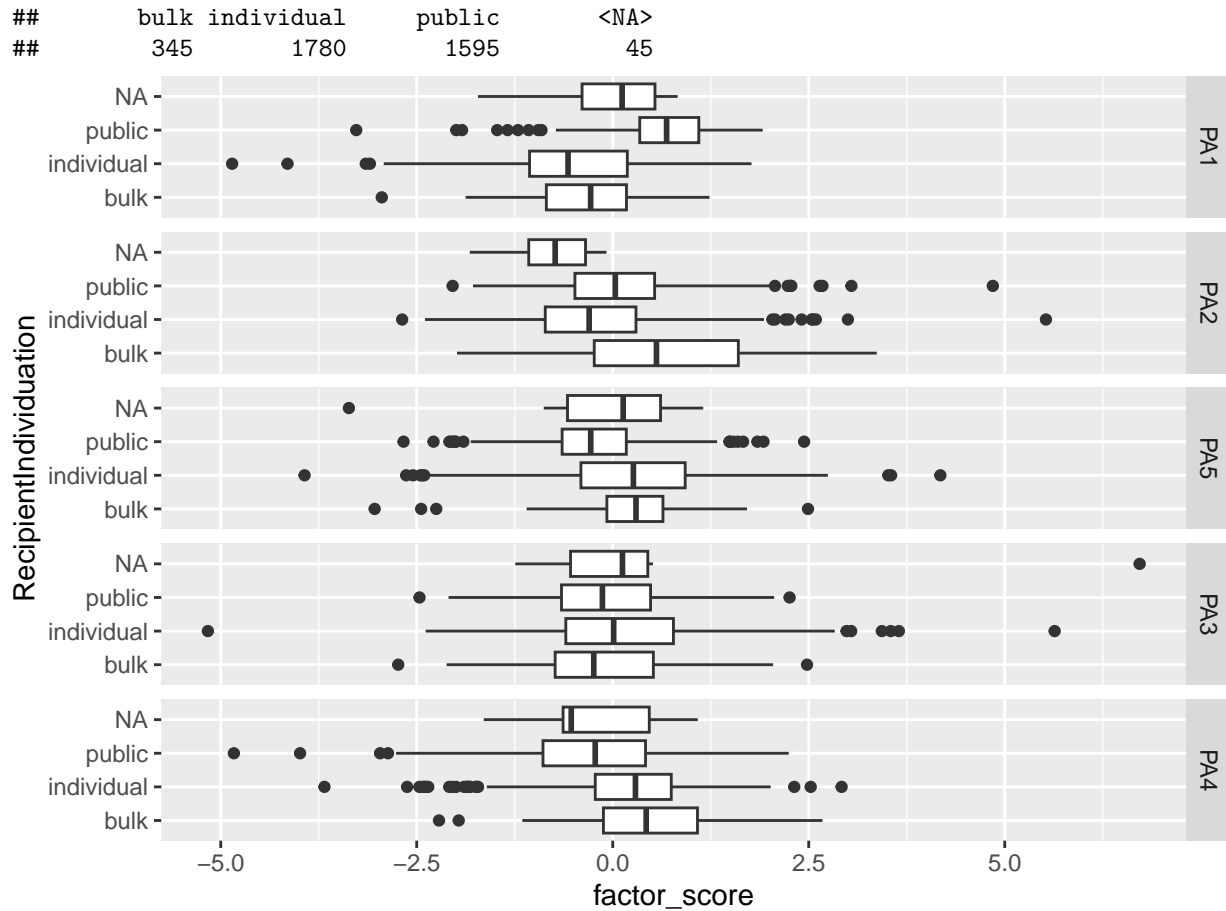
```

court decisions often with RecipientType = combined.

RecipientIndividuation

```
analyze_distributions(res_data$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 = 277.1505, df = 2, p-value = 0
##
##
##           Comparison of x by group
##           (Bonferroni)
## Col Mean-|
## Row Mean |      bulk  individu
## -----+-----
## individu |    1.168717
##           |    0.7276
##           |
## public   |   -8.295974  -16.28068
##           |    0.0000*   0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.369
##
## Test for the significance of differences in RecipientIndividuation over PA2 :
```

```

##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 37.1594, df = 2, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      bulk   individu
## -----+-----
## individu |    5.459269
##          |    0.0000*
##          |
##   public |    3.017217   -4.118228
##          |    0.0077*    0.0001*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0494
##
## Test for the significance of differences in RecipientIndividuation over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 54.4264, df = 2, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      bulk   individu
## -----+-----
## individu |    0.900020
##          |    1.0000
##          |
##   public |    4.832439    6.786504
##          |    0.0000*    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0724
##
## Test for the significance of differences in RecipientIndividuation over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 7.757, df = 2, p-value = 0.02
##
##
##                               Comparison of x by group

```

```

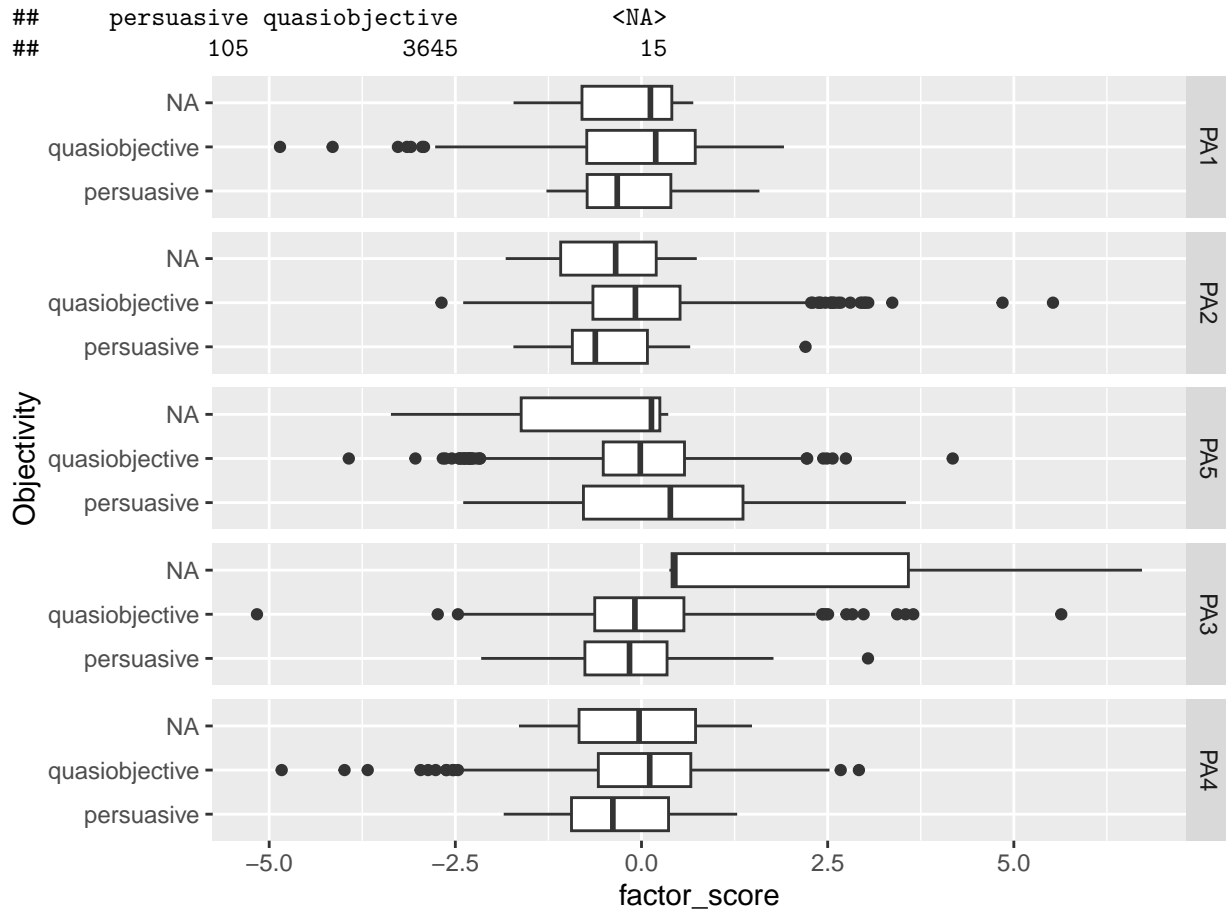
##                                     (Bonferroni)
## Col Mean-|
## Row Mean |      bulk    individu
## -----+-----
## individu |   -1.752331
##          |      0.2392
##          |
##   public |   -0.246344    2.565478
##          |      1.0000    0.0309*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0103
##
## Test for the significance of differences in RecipientIndividuation over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 55.8219, df = 2, p-value = 0
##
##
##                                     Comparison of x by group
##                                     (Bonferroni)
## Col Mean-|
## Row Mean |      bulk    individu
## -----+-----
## individu |    1.625961
##          |      0.3119
##          |
##   public |    5.418144    6.556607
##          |    0.0000*    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0742
##
##   factor kruskal_p epsilon2
## 1   PA1  6.57e-61  0.3690
## 2   PA2  8.53e-09  0.0494
## 3   PA5  1.52e-12  0.0724
## 4   PA3  2.07e-02  0.0103
## 5   PA4  7.56e-13  0.0742
##
## p < 5e-2 found in: PA1 PA2 PA5 PA3 PA4
## p < 1e-2 found in: PA1 PA2 PA5 PA4
## p < 1e-3 found in: PA1 PA2 PA5 PA4
## p < 1e-4 found in: PA1 PA2 PA5 PA4

```

Objectivity

```
analyze_distributions(res_data$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 = 1.5042, df = 1, p-value = 0.22
##
##
##           Comparison of x by group
##           (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## -----+-----
## quasiobj | -1.226477
##           |    0.2200
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.002
##
## Test for the significance of differences in Objectivity over PA2 :
##
##   Kruskal-Wallis rank sum test
##
```

```

## data: x and group
## Kruskal-Wallis chi-squared = 5.7273, df = 1, p-value = 0.02
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## -----+-----
## quasiobj |  -2.393188
##           |    0.0167*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.00762
##
## Test for the significance of differences in Objectivity over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.971, df = 1, p-value = 0.32
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## -----+-----
## quasiobj |   0.985370
##           |    0.3244
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.00129
##
## Test for the significance of differences in Objectivity over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.237, df = 1, p-value = 0.63
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## -----+-----
## quasiobj |  -0.486810
##           |    0.6264
##
## alpha = 0.05
## Reject Ho if p <= alpha

```

```

## epsilon2 = 0.000315
##
## Test for the significance of differences in Objectivity over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 1.6901, df = 1, p-value = 0.19
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |   persuasi
## -----+-----
## quasiobj | -1.300034
##          |    0.1936
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.00225
##
##   factor kruskal_p epsilon2
## 1    PA1    0.2200 0.002000
## 2    PA2    0.0167 0.007620
## 3    PA5    0.3244 0.001290
## 4    PA3    0.6264 0.000315
## 5    PA4    0.1936 0.002250
##
## p < 5e-2 found in: PA2
## p < 1e-2 found in:
## p < 1e-3 found in:
## p < 1e-4 found in:

```

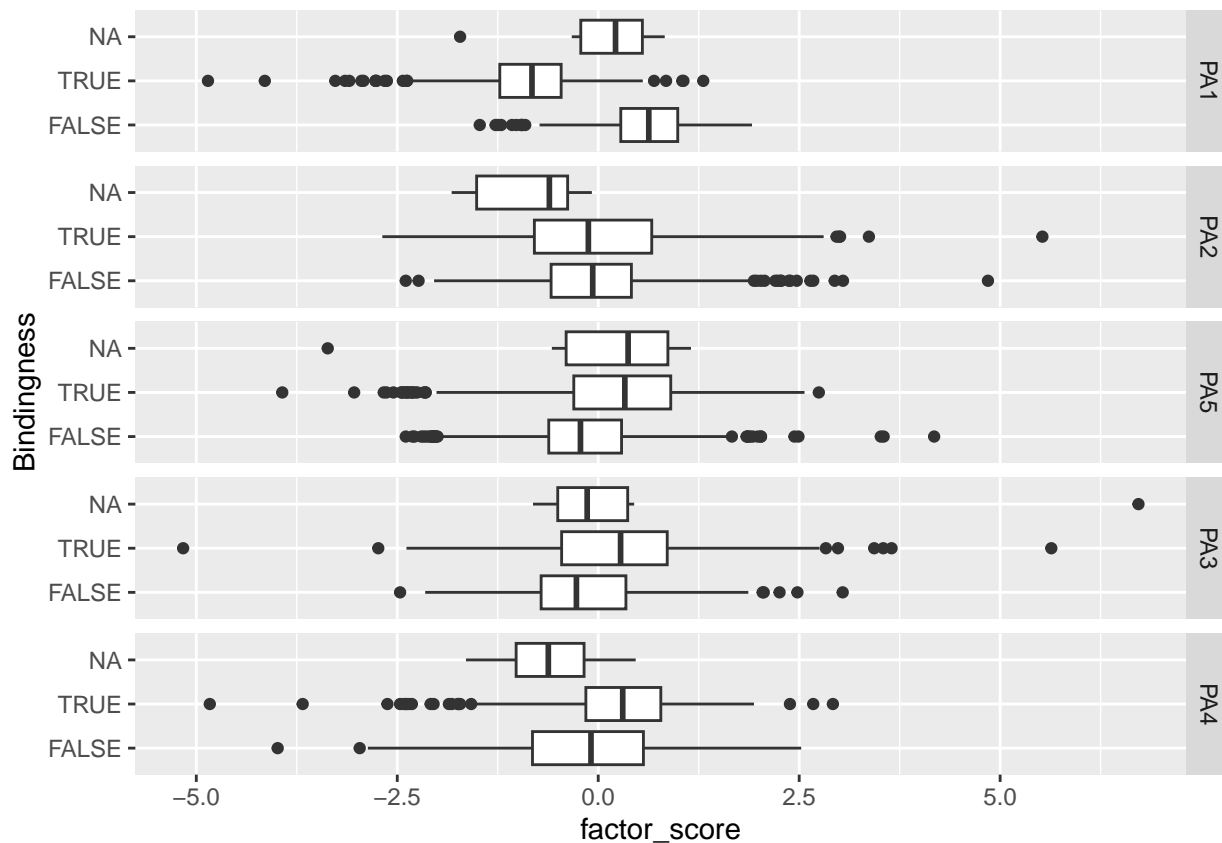
Bindingness

```
analyze_distributions(res_data$long, "Bindingness")
```

```

##
## FALSE TRUE <NA>
## 2220 1515 30

```

```
##
## Test for the significance of differences in Bindingness over PA1 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 438.3294, df = 1, p-value = 0
##
##
##               Comparison of x by group
##               (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## -----+-----
##   TRUE |    20.93631
##       |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.583
##
## Test for the significance of differences in Bindingness over PA2 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 0.0259, df = 1, p-value = 0.87
```

```

##
##
##           Comparison of x by group
##           (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## -----+-----
##      TRUE |    0.160921
##           |    0.8722
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 3.44e-05
##
## Test for the significance of differences in Bindingness over PA5 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 50.2905, df = 1, p-value = 0
##
##
##           Comparison of x by group
##           (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## -----+-----
##      TRUE |   -7.091578
##           |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0669
##
## Test for the significance of differences in Bindingness over PA3 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 39.8349, df = 1, p-value = 0
##
##
##           Comparison of x by group
##           (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## -----+-----
##      TRUE |   -6.311490
##           |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.053
##

```

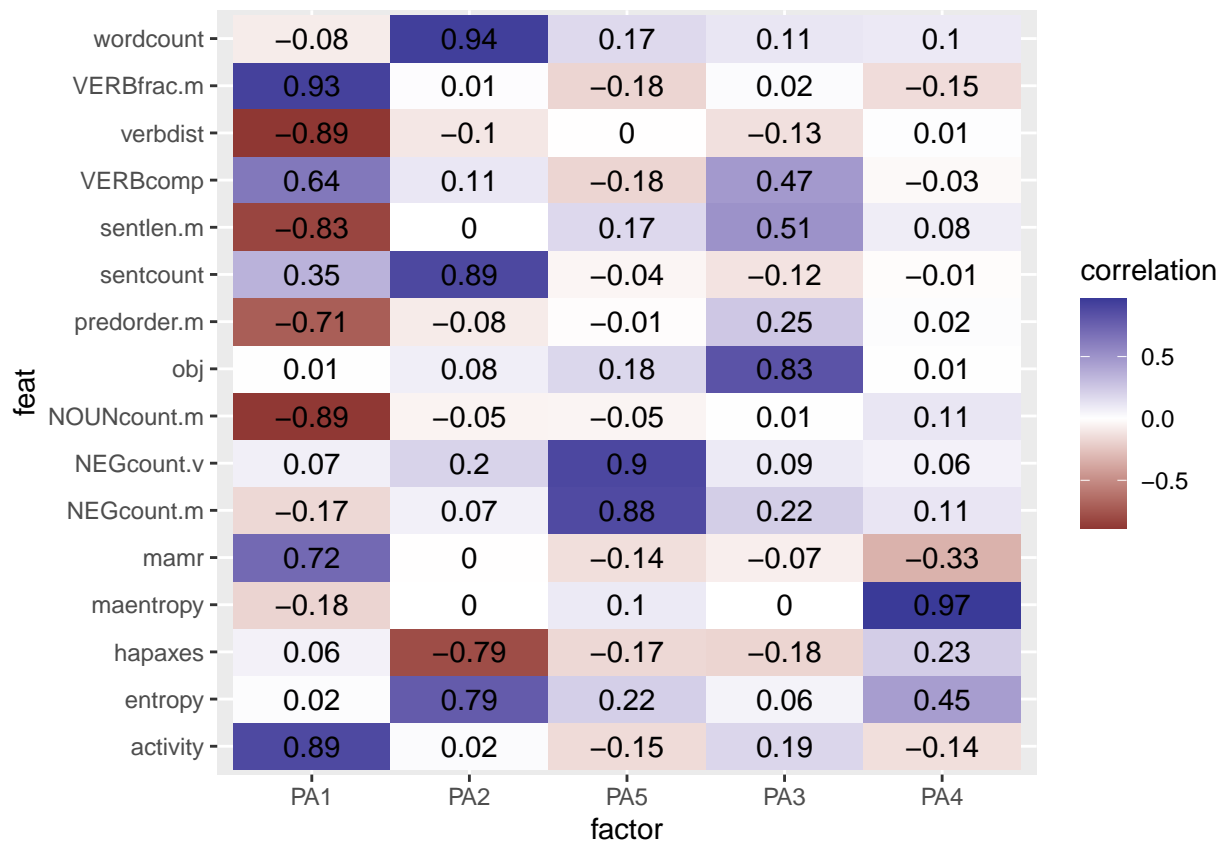
```
## Test for the significance of differences in Bindingness over PA4 :
##
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 29.0874, df = 1, p-value = 0
##
##
##                               Comparison of x by group
##                               (Bonferroni)
## Col Mean-|
## Row Mean |      FALSE
## -----+-----
##      TRUE |  -5.393273
##           |    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha
## epsilon2 = 0.0387
##
##   factor kruskal_p epsilon2
## 1   PA1  2.50e-97 5.83e-01
## 2   PA2  8.72e-01 3.44e-05
## 3   PA5  1.33e-12 6.69e-02
## 4   PA3  2.76e-10 5.30e-02
## 5   PA4  6.92e-08 3.87e-02
##
## p < 5e-2 found in: PA1 PA5 PA3 PA4
## p < 1e-2 found in: PA1 PA5 PA3 PA4
## p < 1e-3 found in: PA1 PA5 PA3 PA4
## p < 1e-4 found in: PA1 PA5 PA3 PA4
```

Feature-factor correlations

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

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

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



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

