ABCD Study CT and Demographic Data Exploratory Data Analysis

Aidan Neher

2023-03-01

What is Exploratory Data Analysis?

Exploratory Data Analysis (EDA) is a structured approach for understanding your data that can be used for research question and hypothesis development. EDA's overall objective is to get insights to make better decisions. Sub-objectives include:

- Identify correlated variables.
- Identify and deal with outliers.
- Identify trends across time.
- Identify trends across space.
- Uncover patterns related to the response variable of interest.
- Create research questions to explore or hypotheses to test.
- Identify possible new data sources.

Set-Up Environment

The .RDS file loaded below was generated using the script "code/0_get_data.R".

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

About the Variables

- subjectkey is the subject's unique identifier.
- eventname is the data collection point for an observation (row of data). Note, interview_age is also available in months.
- Brain structure metrics cortical thickness (thick) and surface area (area) are included. For more on the meaning of these metrics, see https://doi-org.ezp2.lib.umn.edu/10.1007%2Fs00429-015-1177-6

First, we will split our data by eventname to study it cross-sectionally for now.

```
table(tidy_data$eventname)
```

```
##
##    baseline_year_1_arm_1 2_year_follow_up_y_arm_1
##                    11760                    7827
```

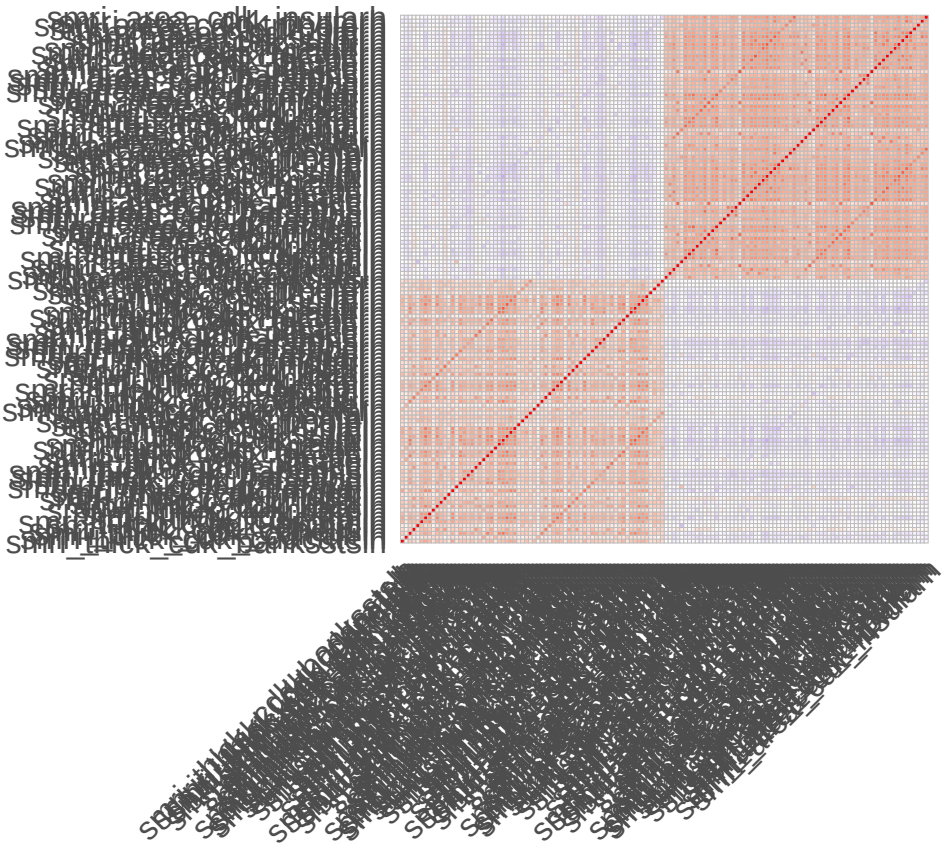
```
split_data = split(tidy_data, f = tidy_data$eventname)
baseline_data = split_data$baseline_year_1_arm_1 %>% ungroup
baseline_smri = select(baseline_data, starts_with("smri"))
```

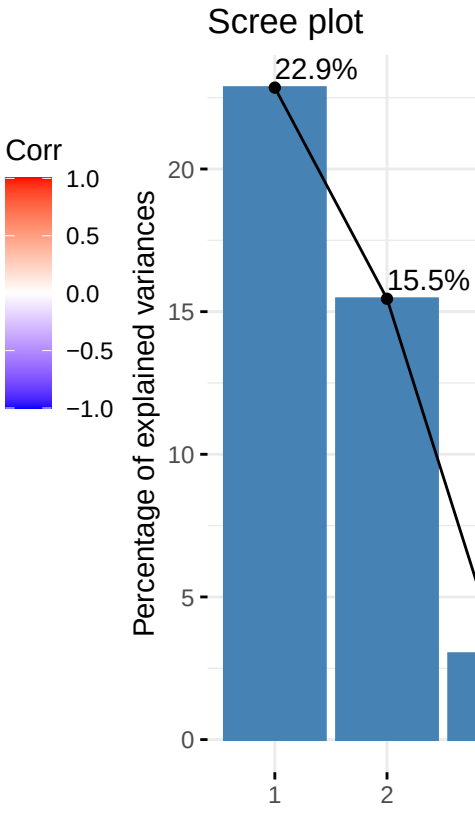
Principal Components Analysis (PCA)

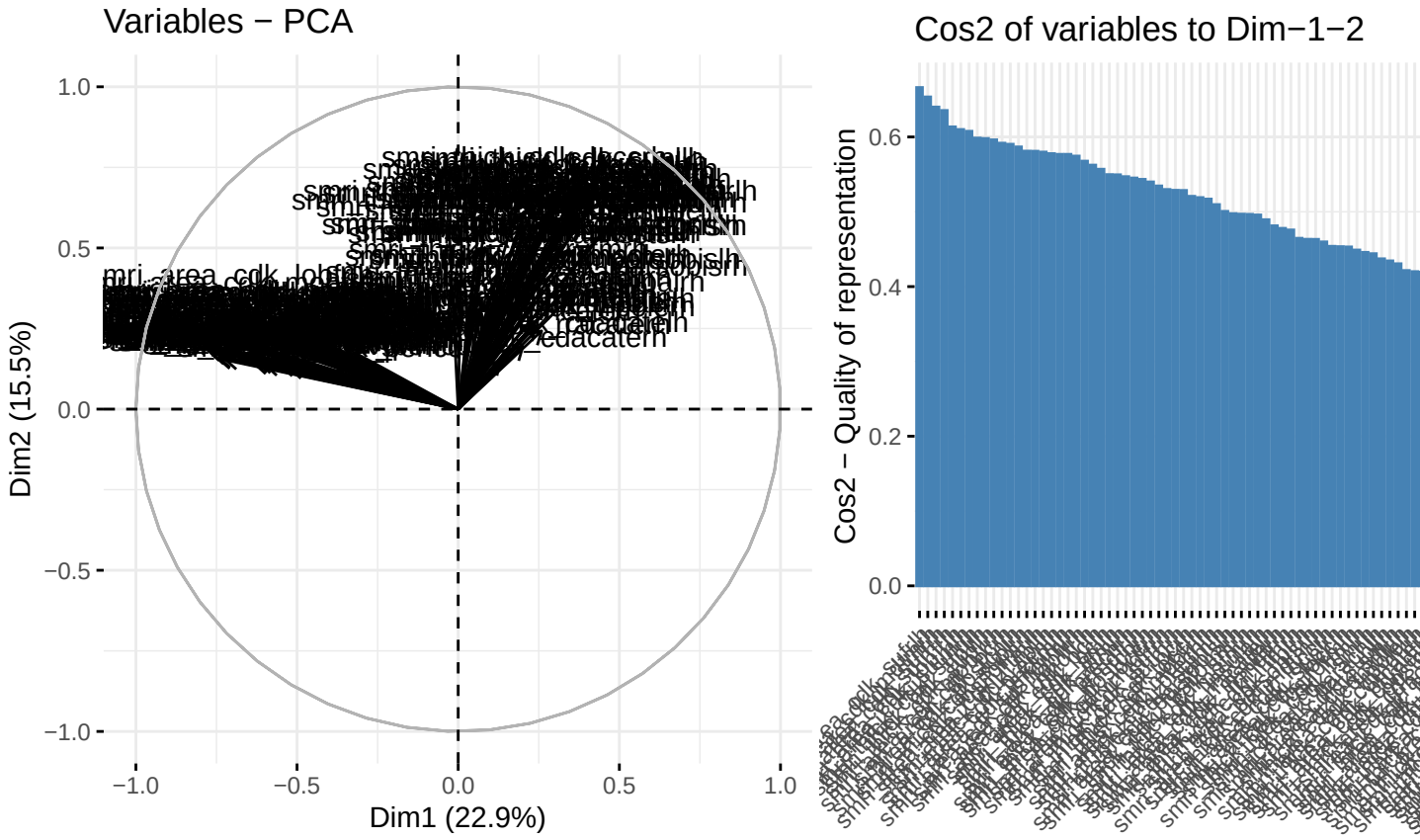
For Principal Components Analysis (PCA), the R function `prcomp()` is preferred. Note, the loadings are accessible in the resulting object's rotation feature.

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

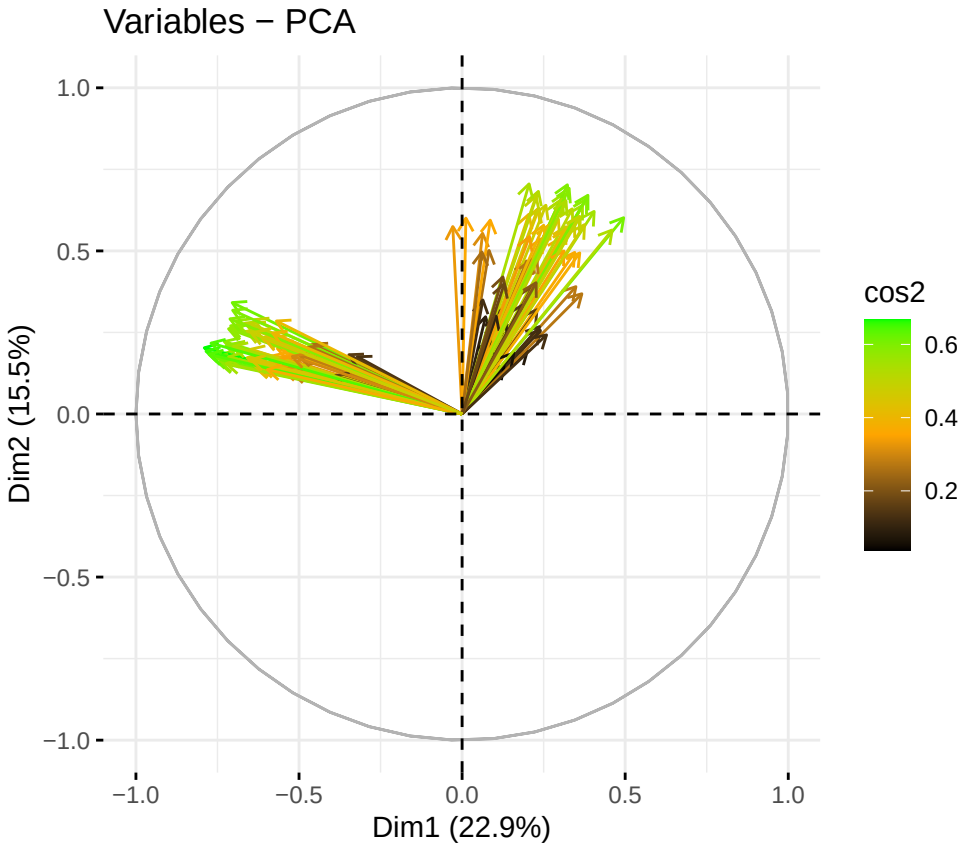
```
## [1] 0
```








```
## Warning: ggrepel: 136 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Principal Components Regression (PCR)

Next we perform regression and classification against a clinical severity score and a binary clinical outcome respectively using principal components identified in the PCA above and covariates.

```
# TODO: use lmer to account for fixed/ random effects, control for site effect and family within site
library(lme4)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
outcome_of_interest <- "outcome_internalizing_score"
outcome_names <- c("outcome_si", "outcome_internalizing_score")
outcome_to_remove <- subset(outcome_names, outcome_names!=outcome_of_interest)

base_model_data <- baseline_data %>%
  select(-starts_with("smri")) %>%
  # choose and rename outcome of interest
  select(-c("subjectkey", "eventname", all_of(outcome_to_remove))) %>%
  rename(outcome = starts_with("outcome"))

var_names <- colnames(base_model_data)
```

```r
random_effect_index <- which(var_names  %in% c("abcd_site", "rel_family_id"))
outcome_index <- which(var_names == "outcome")

base_fixed_effects <- var_names[-c(random_effect_index, outcome_index)] %>% paste(collapse = "+")
base_formula <- paste0("outcome~", base_fixed_effects, "+(1|abcd_site/rel_family_id)")


# TODO: https://stats.stackexchange.com/questions/22988/how-to-obtain-the-p-value-check-significance-of

# compute a model where the effect of PC is not estimated
restricted_fit = lmer(
  data = base_model_data,
  formula = base_formula,
  REML = F #because we want to compare models on likelihood
)

fits <- list(restricted_fit)

for (i in 1:10) {
  pc_index <- seq(1, i)
  if (i==1) {
    model_data <- base_model_data %>%
      cbind(PC1=pca$x[,pc_index]) # TODO: Find a substitute for this if else
  } else {
    model_data <- base_model_data %>%
      cbind(pca$x[,pc_index])
  }
  pc_names <- paste0("PC", pc_index, collapse = "+")
  model_formula <- paste(base_formula, pc_names, sep = "+")
  # compute a model where the effect of an additional PC is estimated
  fits[[i+1]] = lmer(
    data = model_data,
    formula = model_formula,
    REML = F #because we want to compare models on likelihood
  )
}
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00427516 (tol = 0.002, component 1)
```

```r
likelihood_ratios <- list()
for (i in 1:10) {
  unrestricted_fit <- fits[[i]]
  restricted_fit <- fits[[i+1]]
  # compute the AIC-corrected log-base-2 likelihood ratio (a.k.a. "bits" of evidence)
  likelihood_ratios[[i]] <- (AIC(restricted_fit)-AIC(unrestricted_fit))*log2(exp(1))
}

likelihood_ratios
```

```
## [[1]]
## [1] -14.76303
##
## [[2]]
## [1] -6.020721
```

```
##
## [[3]]
## [1] -2.868523
##
## [[4]]
## [1] 0.5724099
##
## [[5]]
## [1] 2.856471
##
## [[6]]
## [1] 2.099211
##
## [[7]]
## [1] 2.820005
##
## [[8]]
## [1] 2.606116
##
## [[9]]
## [1] -0.07920071
##
## [[10]]
## [1] 2.703107
```

```
# classification_data <- model_data %>%
#   select(-c("subjectkey","eventname","outcome_internalizing_score")) %>%
#   rename(outcome = outcome_si)
# classification_fit <- glmer(model_formula, data = classification_data, family = binomial)
# summary(classification_fit)
```

Principal Component Loading Plots

Now, let's make plots of the first ten principal components' thickness loading minus surface area loading. Thus,

- a region with a more positive value is more represented by cortical thickness,
- a region with a near zero value is represented by both cortical thickness and surface area, and
- a region with a more negative value is more represented by surface area.

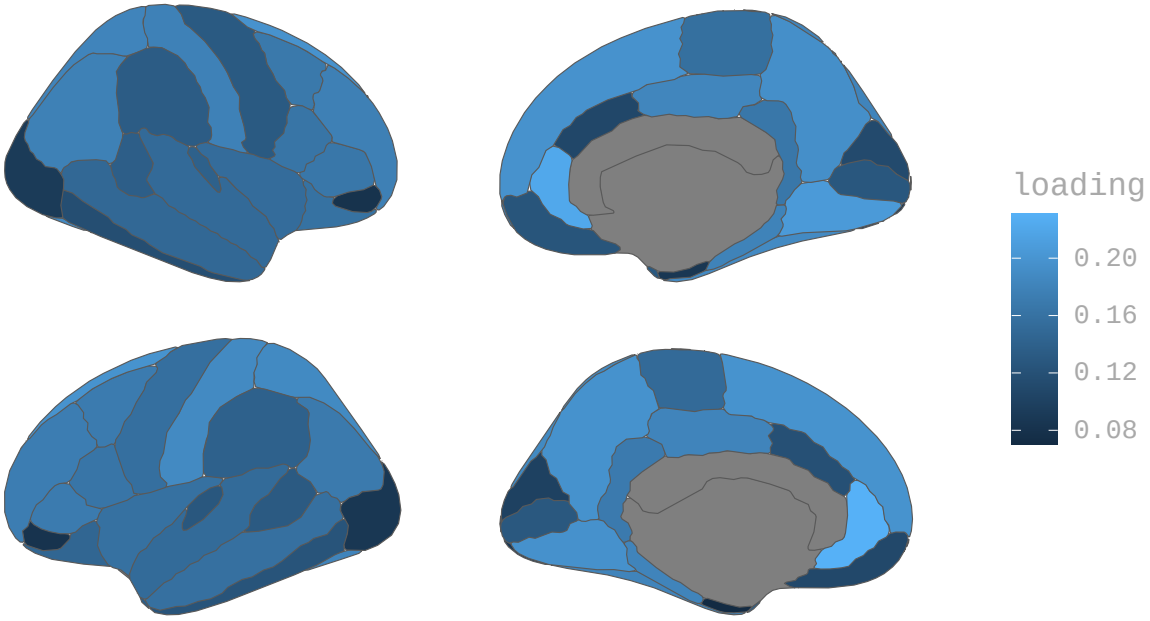
Consider, Is a given PC representing variance in certain regions? Is a given PC more dominated by thickness or surface area?

```
library(ggseg) # https://drmowinckels.io/blog/2021-03-14-new-ggseg-with-geom/

plot_pc_relative_loadings <- function(pc) {
  pc_thick_index <- which(names(pc) %>% str_detect("thick"))
  pc_thick <- pc[pc_thick_index]
  pc_area <- pc[-pc_thick_index]
  regions <- dk$data %>%
    filter(region != "corpus callosum") %>%
    pull(region) %>% na.omit %>% unique %>% sort
  df <- data.frame(hemi=c(rep("left", length(regions)),
                          rep("right", length(regions))),
                   region=rep(regions, 2),
                   loading=pc_thick-pc_area,
```

```
                    row.names = NULL)
  ggplot(df) + geom_brain(atlas = dk,
               position = position_brain(hemi ~ side),
               aes(fill = loading)) + theme_brain2()
}

pc_plots <- apply(pca$rotation[,1:10], 2, plot_pc_relative_loadings)
pc_plots
```

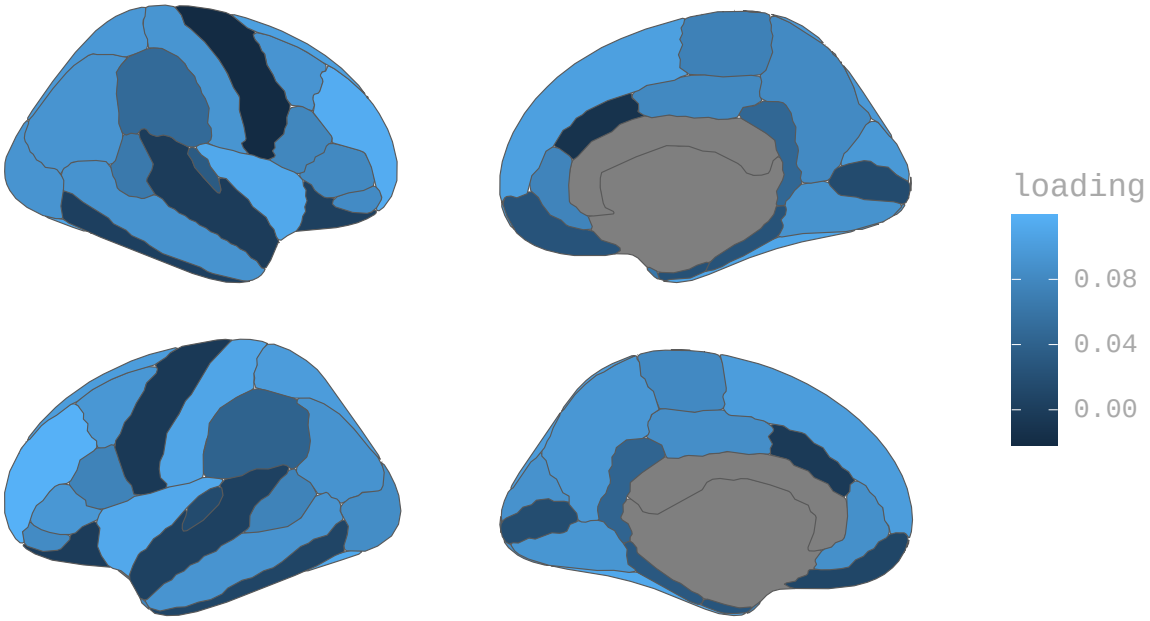
```
## $PC1
```

```
## merging atlas and data by 'hemi', 'region'
```



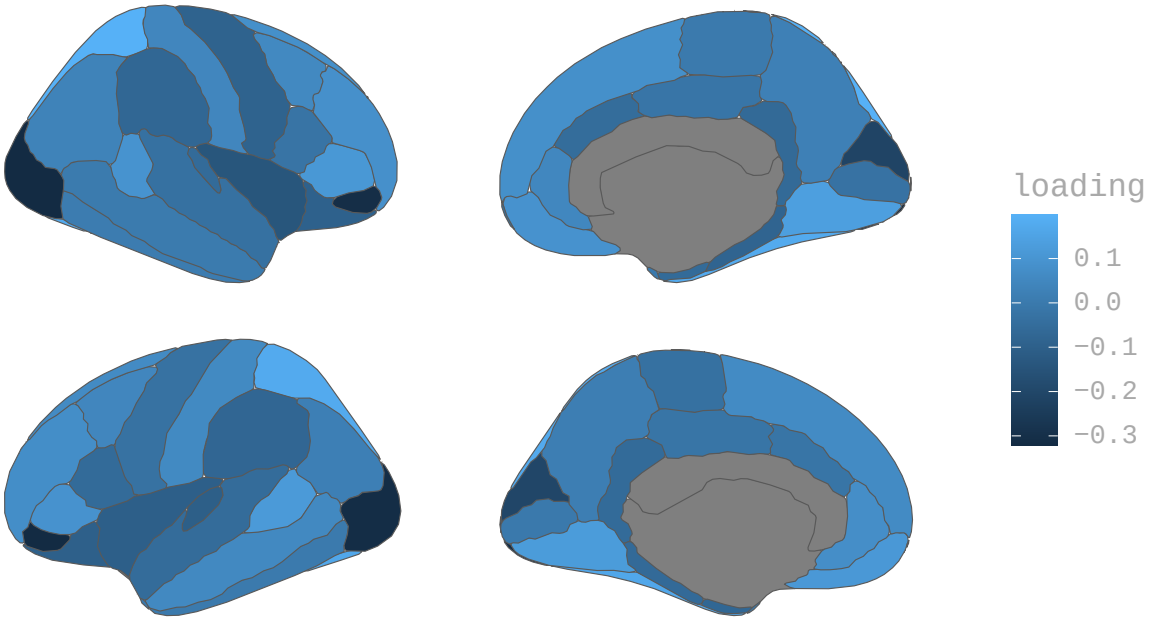

```
##
## $PC2
```

```
## merging atlas and data by 'hemi', 'region'
```



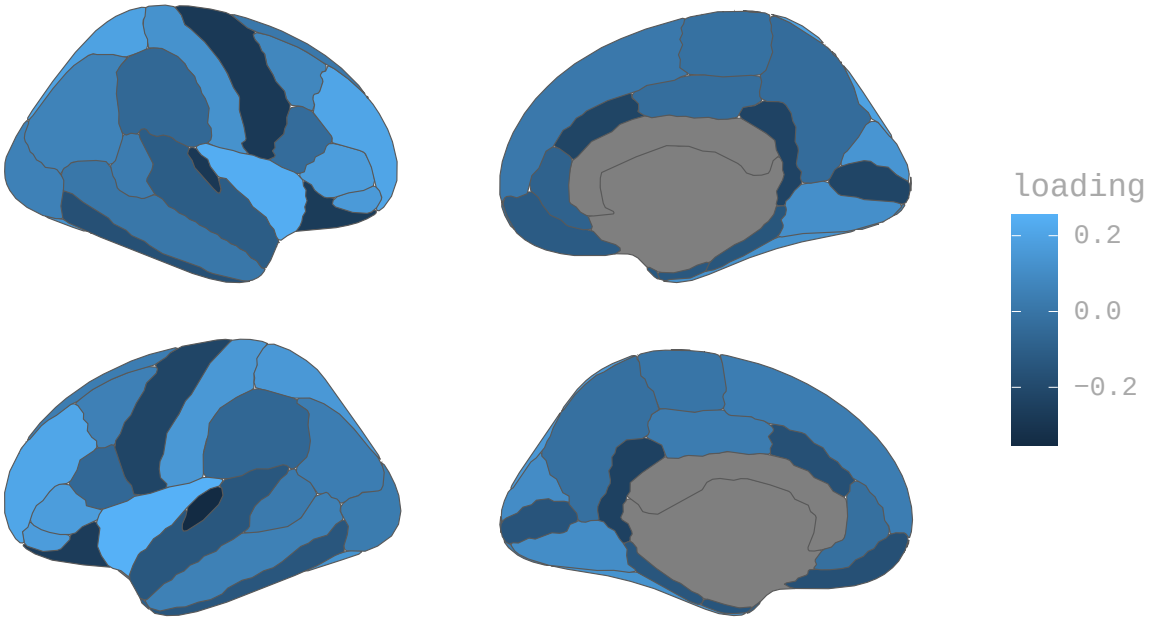

```
##
## $PC3
```

```
## merging atlas and data by 'hemi', 'region'
```



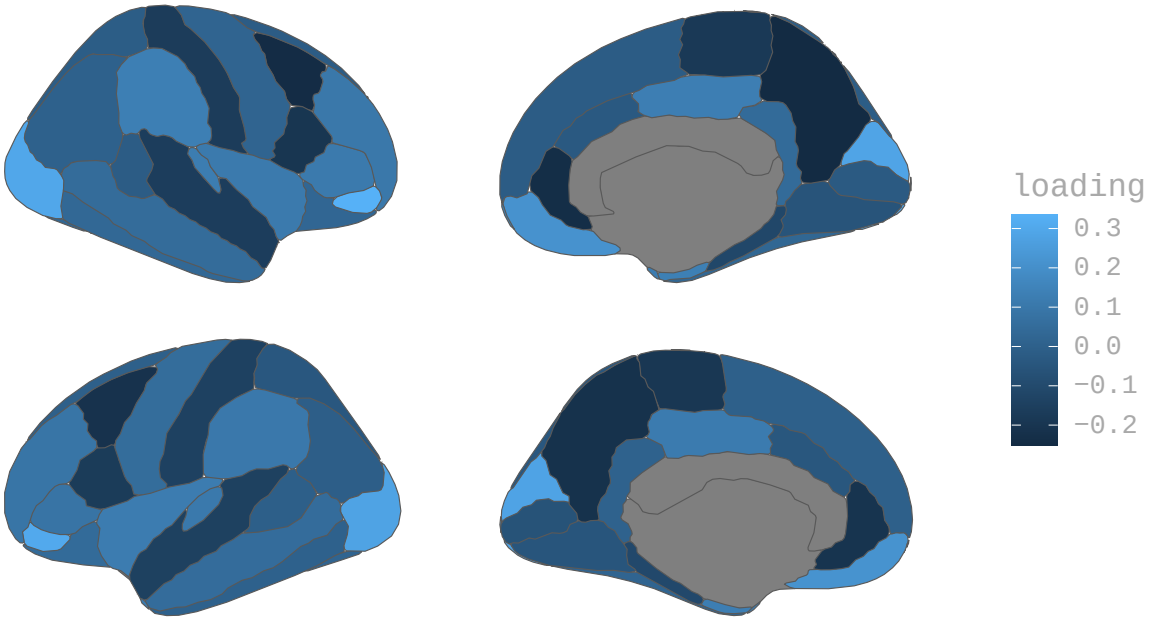

```
##
## $PC4
```

```
## merging atlas and data by 'hemi', 'region'
```



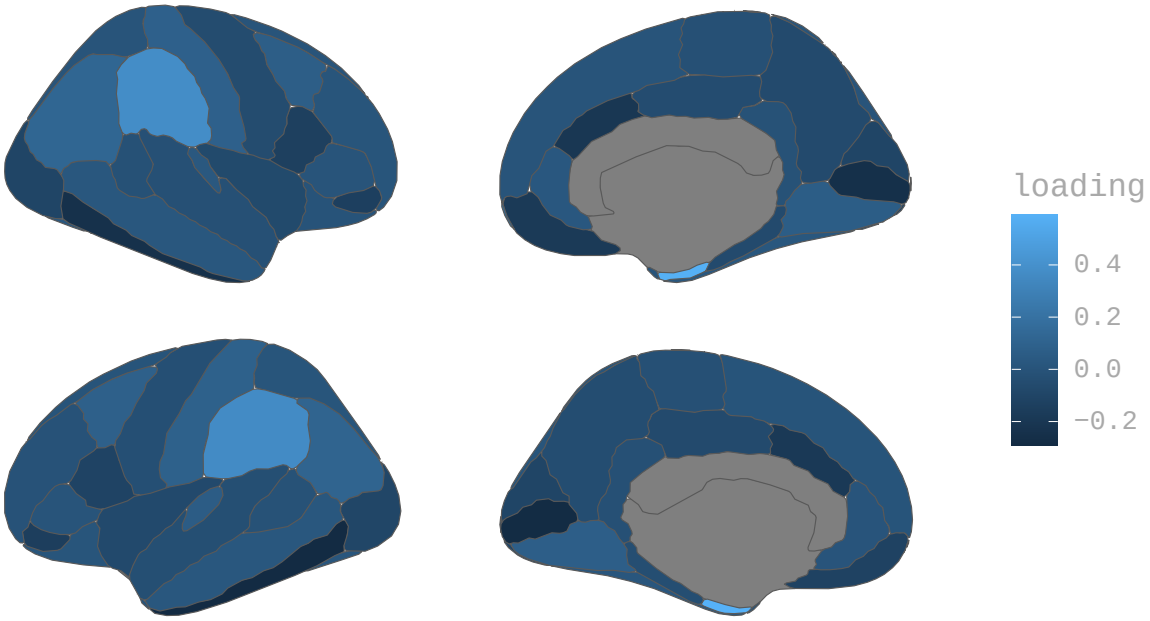

```
##
## $PC5
```

```
## merging atlas and data by 'hemi', 'region'
```



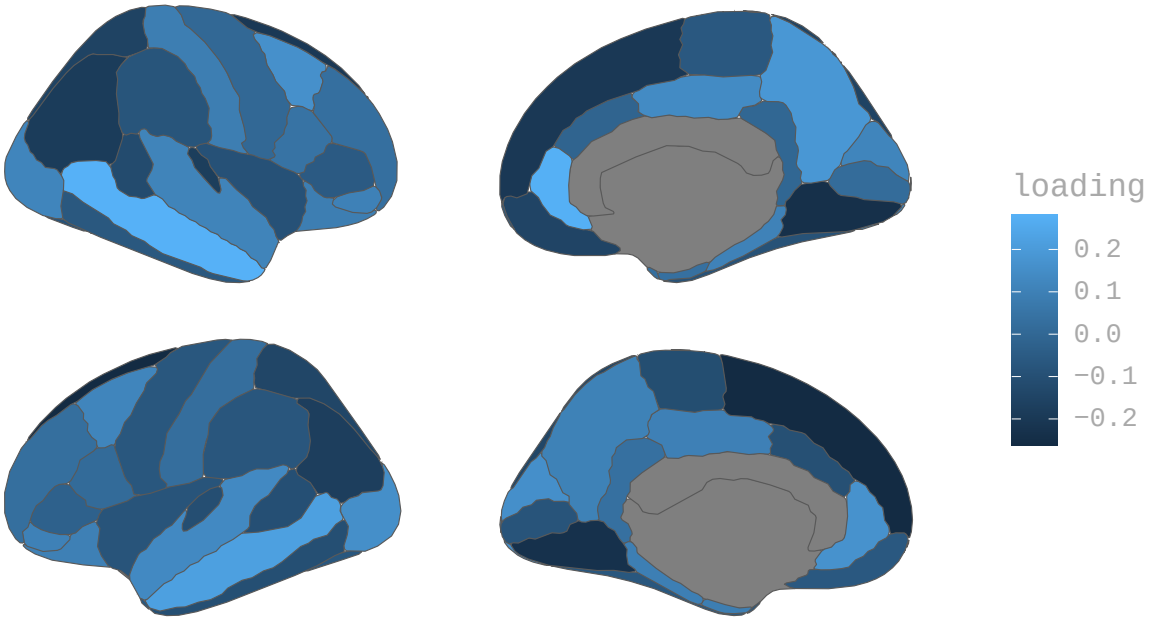

```
##
## $PC6
```

```
## merging atlas and data by 'hemi', 'region'
```



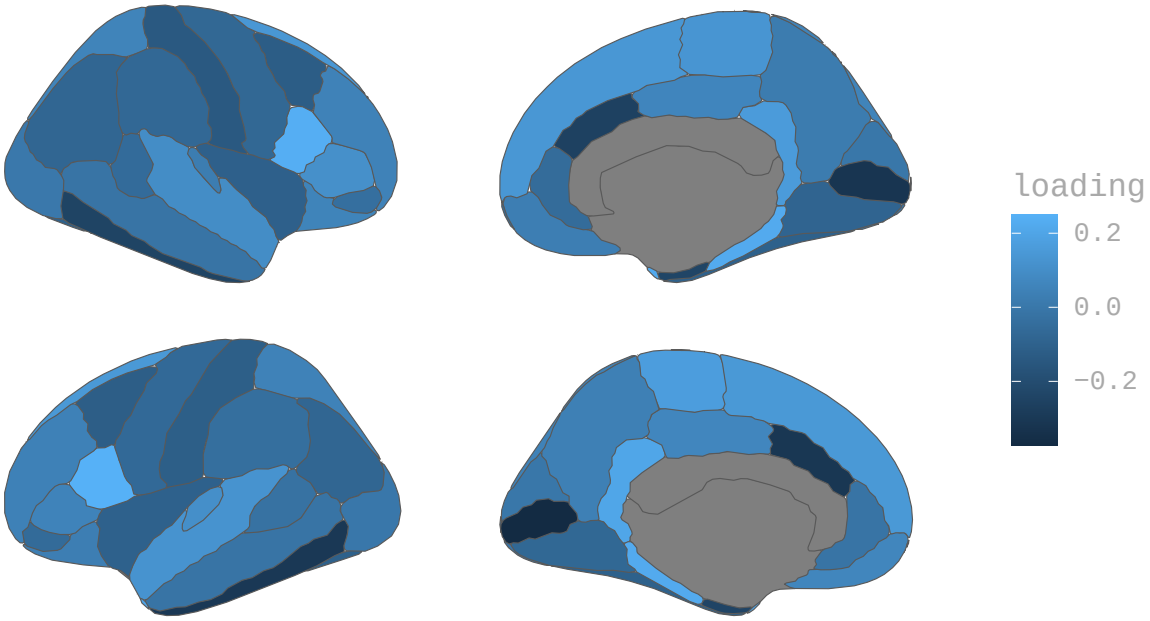

```
##
## $PC7
```

```
## merging atlas and data by 'hemi', 'region'
```



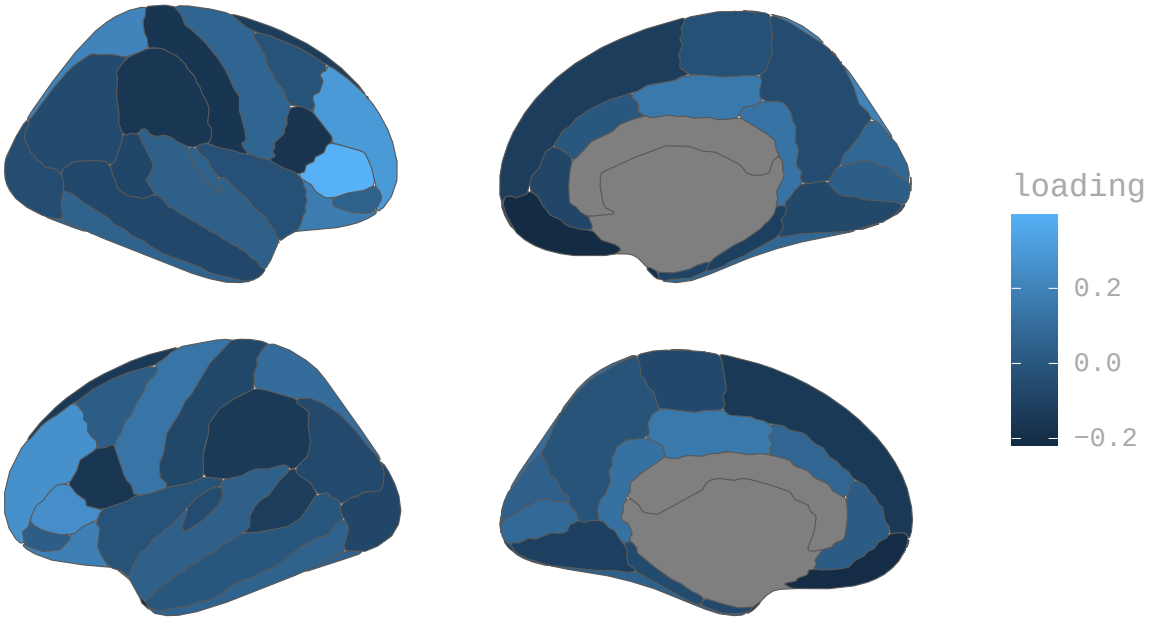

```
##
## $PC8
```

```
## merging atlas and data by 'hemi', 'region'
```



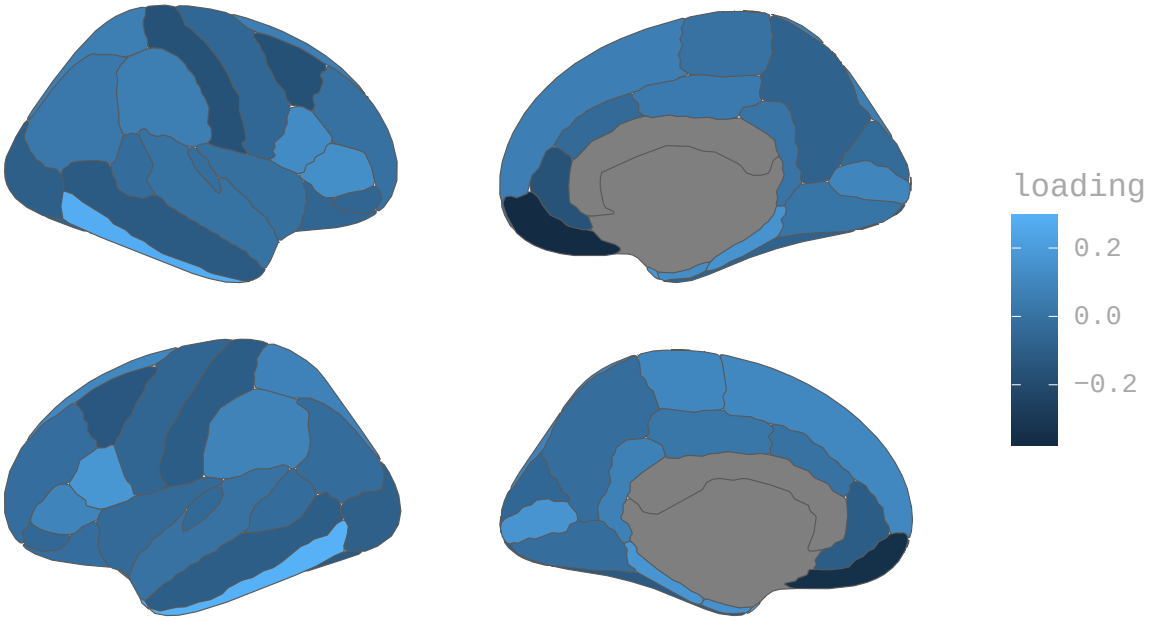

```
##
## $PC9
```

```
## merging atlas and data by 'hemi', 'region'
```




```
##
## $PC10
```

```
## merging atlas and data by 'hemi', 'region'
```



Questions

- What other covariates are required?
- Do we want to use eventname or interview age for temporal effect?