Final Project Data Exploration

Denis Ostroushko

Questions proposed by Erika are in red, my comments and suggestions are in black.

1

Which observations will be included in the analysis? It may be that your population of interest does not include everyone in the original data set you obtained. You should describe any exclusion criteria that you will apply.

Looking for outliers in the data

Figure 1: primary outcome: reading scores: no outliers

Figure 2: primary outcome: reading scores: some outliers that have higher than 'normal' scores in the early periods, then likely these individuals regress to the mean

Figure 3 : yes, there are outliers, which are 'natural', and we will not do anything about them.

2

Will any transformations (e.g., log) be applied to the data prior to carrying out the analysis? Additionally, it should be clear how each variable will be treated in your analysis (e.g. binary, continuous, categorical) and whether each predictor is cluster variant or invariant.

Primary outcome

No transformation required to the primary outcome.

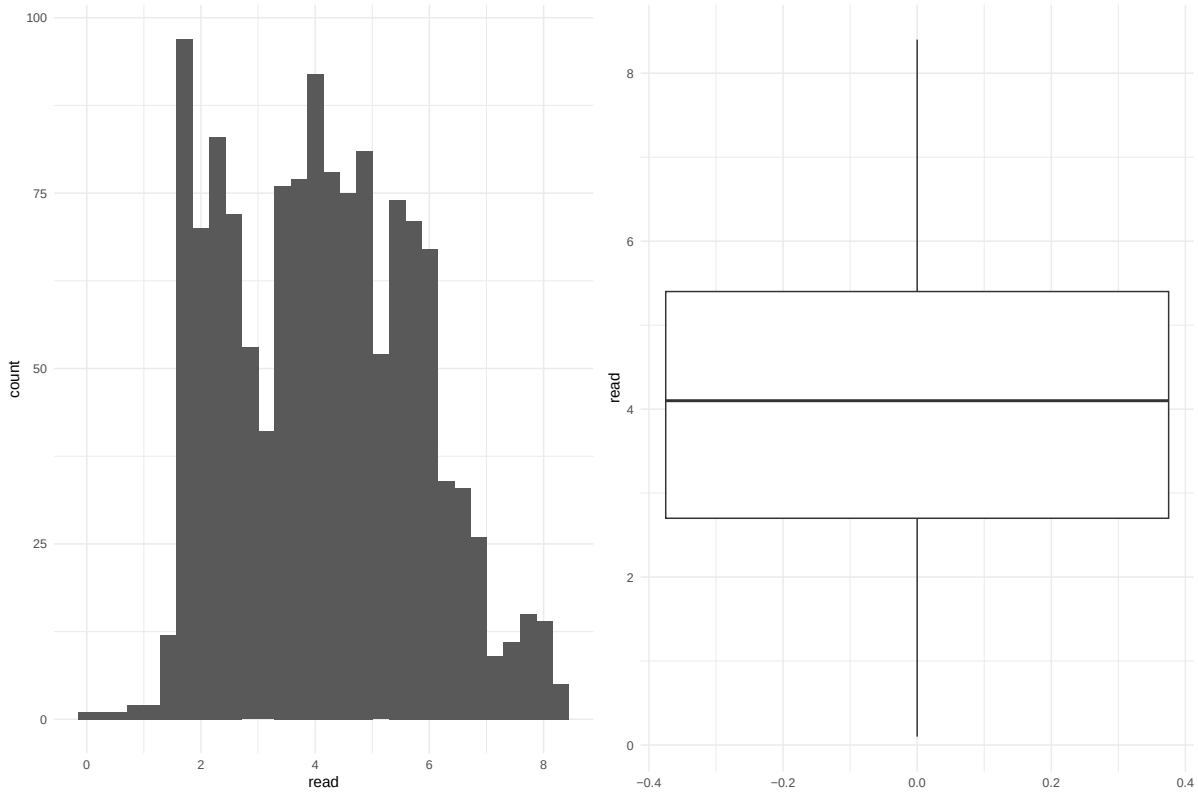


Figure 1: No Outliers in the primary outcome

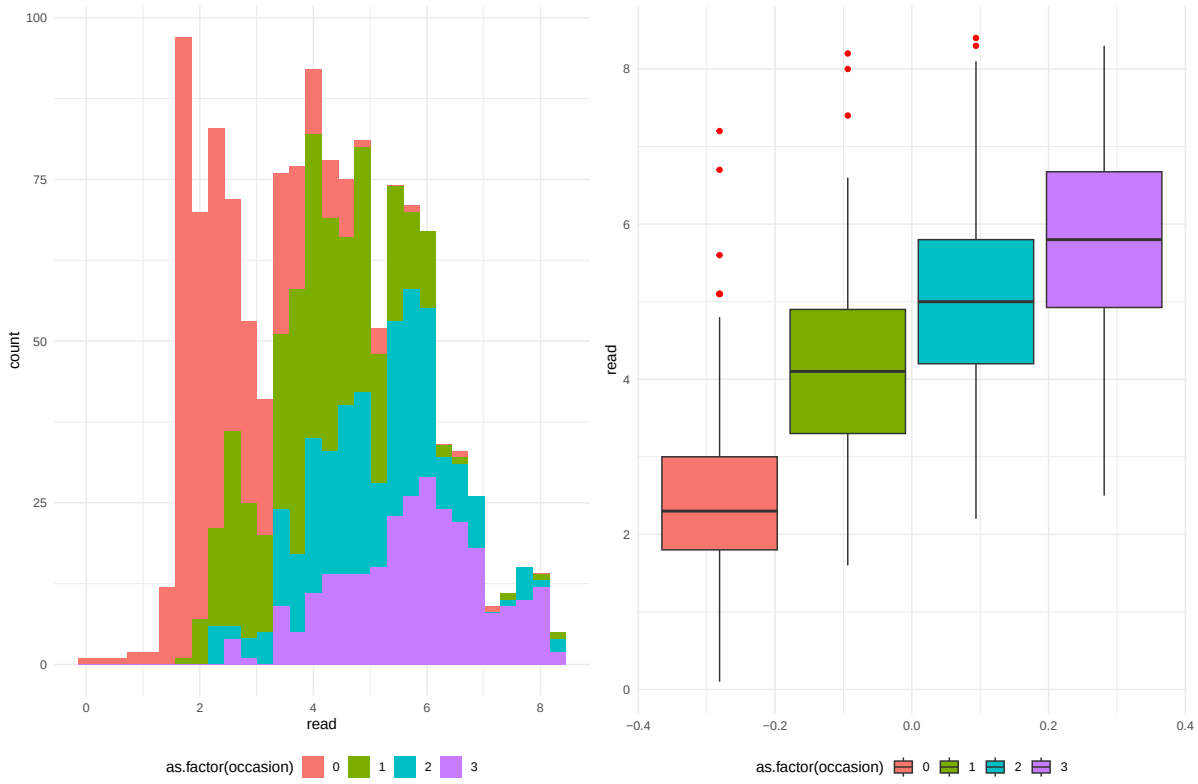


Figure 2: Some outliers in the early observation periods

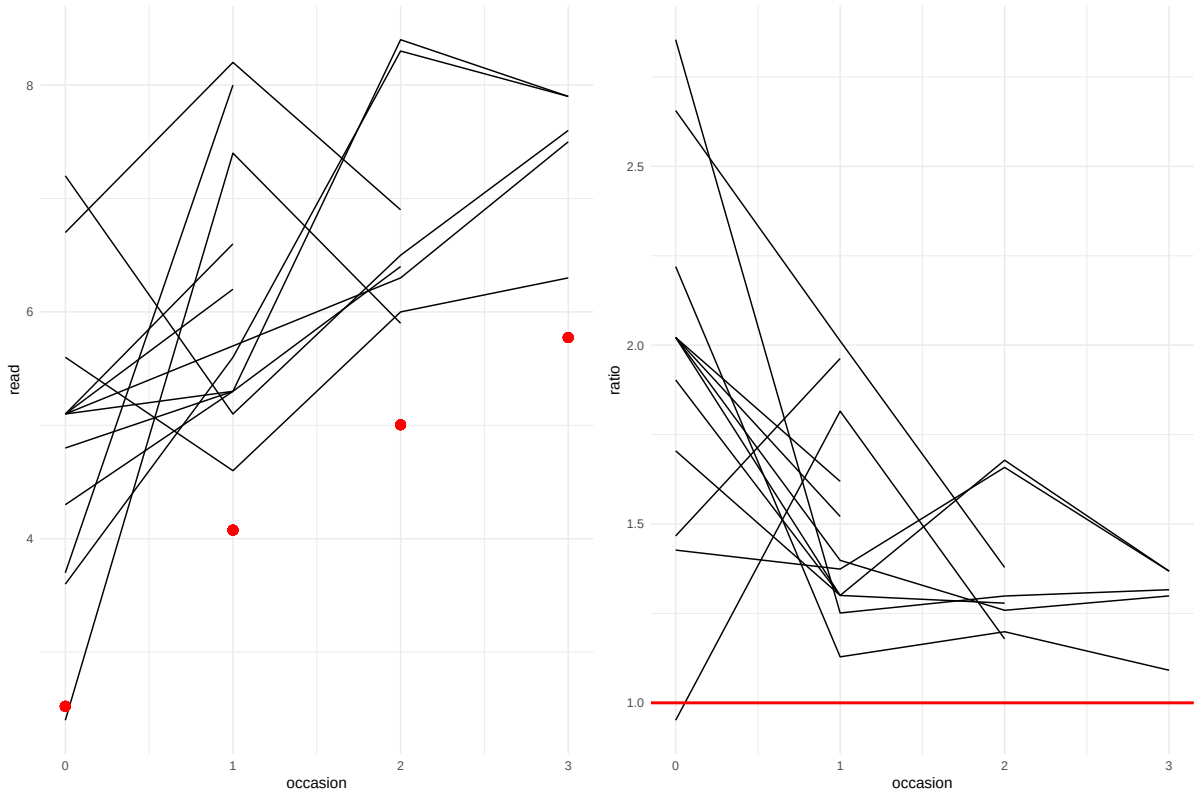


Figure 3: Profile of IDs that are outliers on the boxplot at any point. They regress to the mean

Secondary outcome

Figure 4 : highly skewed outcome. Consider log-10 transformation, since highest value of 10 will become 1, which is convenient. Still, the outcome is skewed.

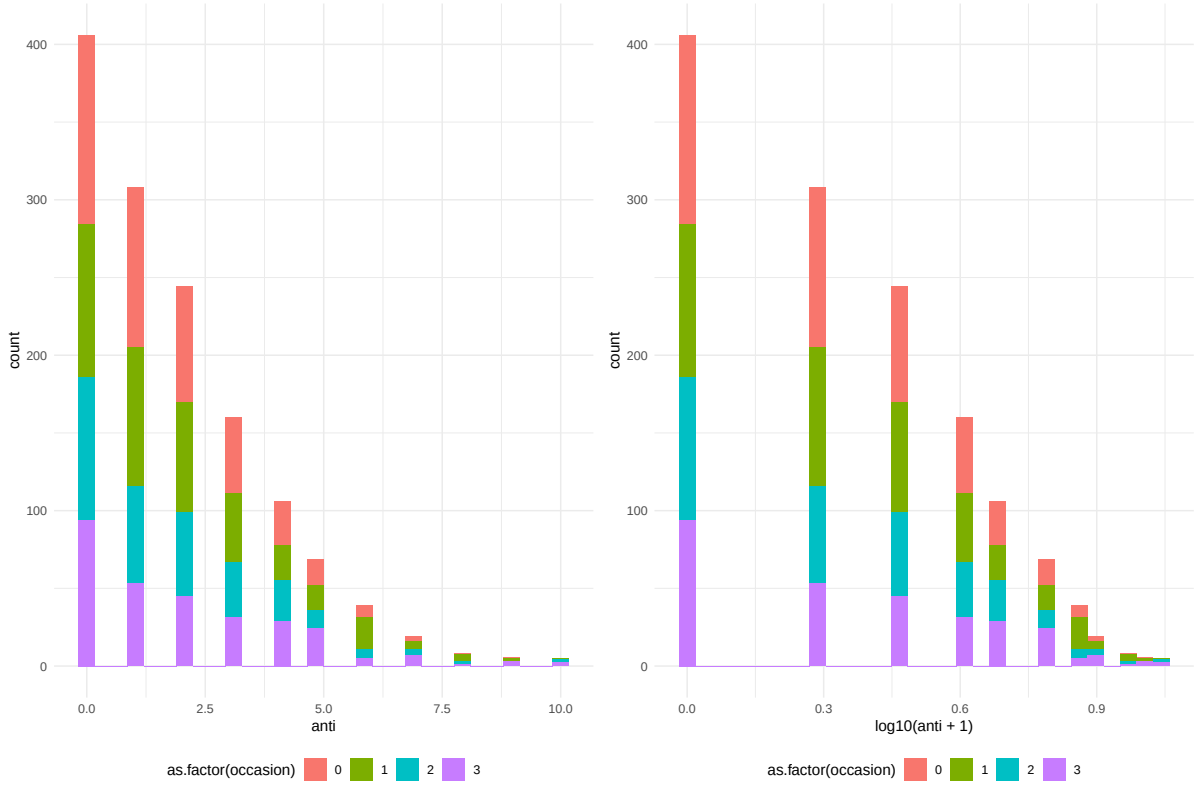


Figure 4: Some outliers in the early observation periods

3

What regression methods will be used? a. Regression type (GLM/GEE/mixed models) b. Family (Gaussian/Binomial/Poisson/Gamma/etc.) c. Link/variance function d. Effects of interest e. Any adjustment variables included

A - Primary Outcome

GEE Model	Random Effects Model
GEE Pro: Will get approximately correct standard errors for effects using sandwich variance estimator	RE Pro: Can account for individual effects, in slope and intercept, which can explain most of variation
GEE Pro: We focus on marginal effects, time-variant effects are not of primary interest to us	RE Con: random effects may account for more variation than fixed effects. Fixed effects are primary focus here
GEE Con: might be less efficient, C.I. could be too wide	RE Pro: IF we indeed have an exchangeable correlation structure, we can have more efficient variance estimates
	RE Con: Random Effects might not be fitting the distributional assumptions well

Details:

Figure 5: trend of reading score against follow p occasion and child age: looks like there is high variance in individual intercepts.

Random slopes: not so clear. Some profiles, in purple, seem to have strong linear trend and low variation within the cluster. *Should we consider random slopes?* Perhaps. This can be a hypothesis to test. It is really hard to determine this due to the large number of individual clusters in the data.

Ultimately, we want to test the effect of a cluster invariant covariate, so GEE might be a little more appropriate. We also have a large number of clusters, 405, so sandwich variance estimator may be useful here to correct for any incorrect assumptions that we make.

A - Secondary Outcome

GEE might not be useful since we just wont detect any significant effects to interpret

On the other hand, a Random Effects Model can show that a between cluster variation accounts for the majority of variation in the data. We will probably will not find any interesting fixed effects, but will be able to show that there is significant amount of variation in the data, and it should be explained different predictors.

B

Outcome is positive values, which are not integers. Some of them are close to zero. We can use Gaussian with log link, or Gamma with log link. Whatever one we select, it is important to use log-link to make sure that our regression model does not have negative fitted values.

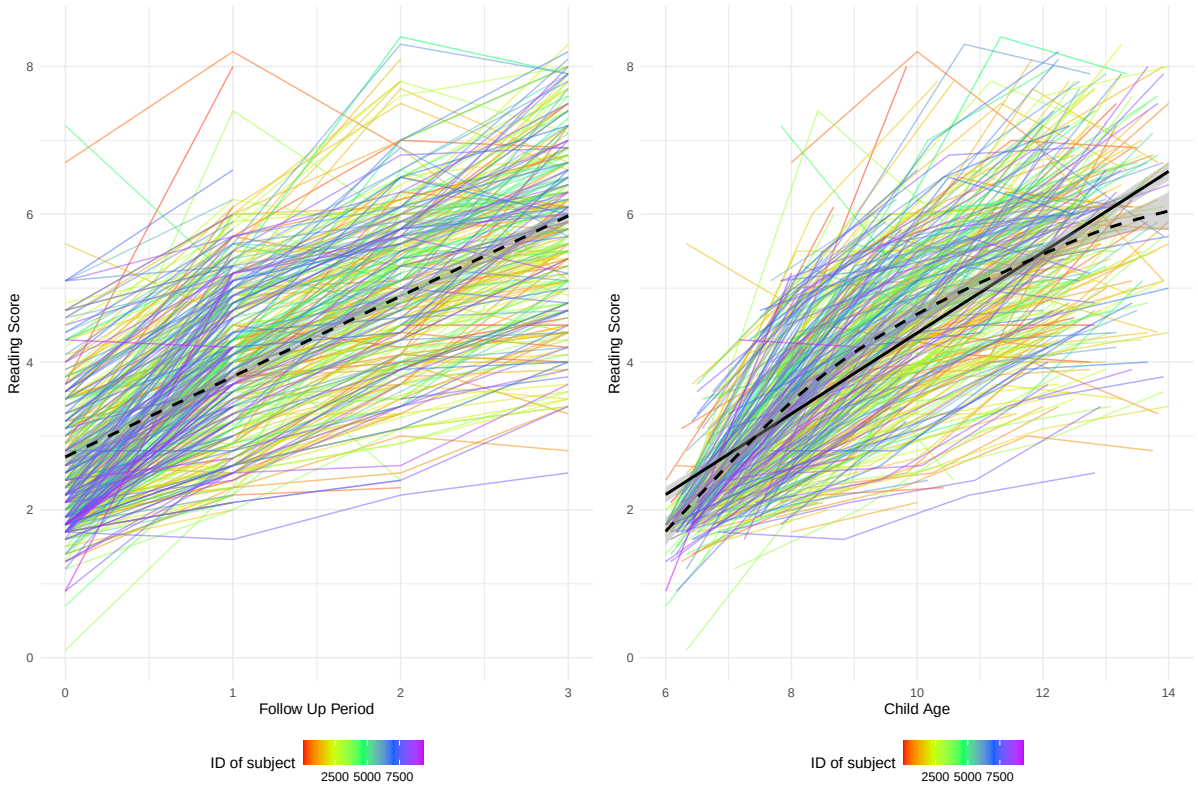


Figure 5: Trend of Reading ability vs Follow Up Time and Kid's age for Individuals

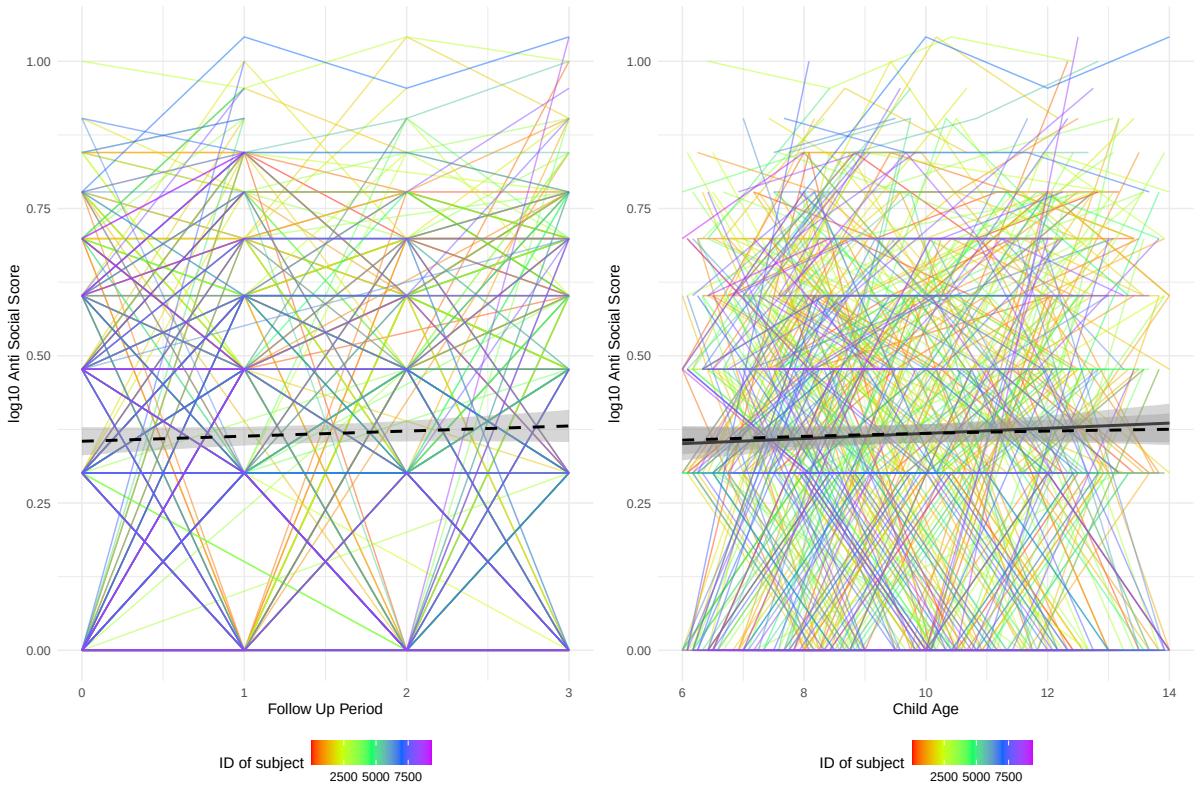


Figure 6: Trend of Reading ability vs Follow Up Time and Kid's age for Individuals

C

see **B**

D

primary effect of interest: effect of the measure of emotional and cognitive stimulation at home on the reading ability. This value is measured at baseline and therefore is a cluster invariant covariate. We will consider their marginal effects as well as the interaction between the two.

As a secondary outcome, we consider anti-social behavior measured using a 0-10 scale. We convert these scores to the log-10 scale, and investigate the effect of at home emotional and cognitive stimulation on the anti-social behavior.

E

Other adjustment variables:

1. Mom's age: speculatively, higher age can be a proxy measure for varying access to parental resources, money, more time to spend with the family due to already advanced stage of a career, etc…
2. Kid's age: clearly older children should have higher levels of reading scores
3. Kid's gender/sex: a binary predictor

4

What assumptions will be made in fitting these models? a. (for GEE) Working correlation(s) b. (for mixed models) Random intercepts/slopes

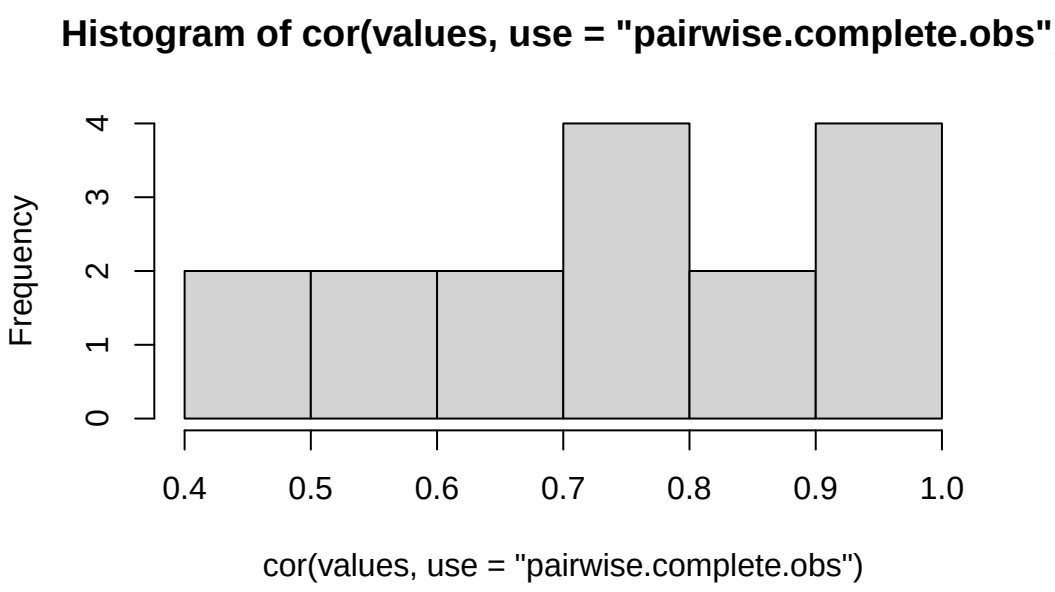
4 - A

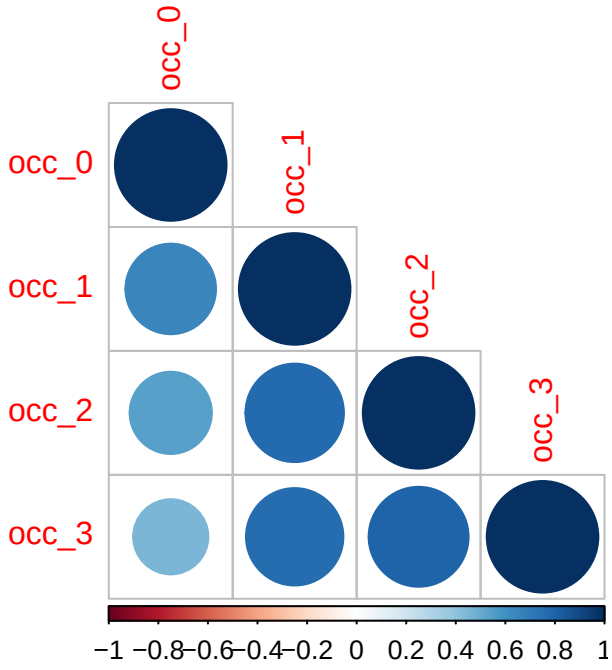
Primary Outcome

Choice of correlation might not be important, since we have a very large sample for sandwich to work its magic.

Based on printed correlation matrix below, pick either AR1 or banded, since observations that are further apart in time are less correlated.

```
          occ_0     occ_1     occ_2     occ_3
occ_0 1.0000000 0.6588808 0.5429116 0.4545546
occ_1 0.6588808 1.0000000 0.7789734 0.7622562
occ_2 0.5429116 0.7789734 1.0000000 0.8019366
occ_3 0.4545546 0.7622562 0.8019366 1.0000000
```

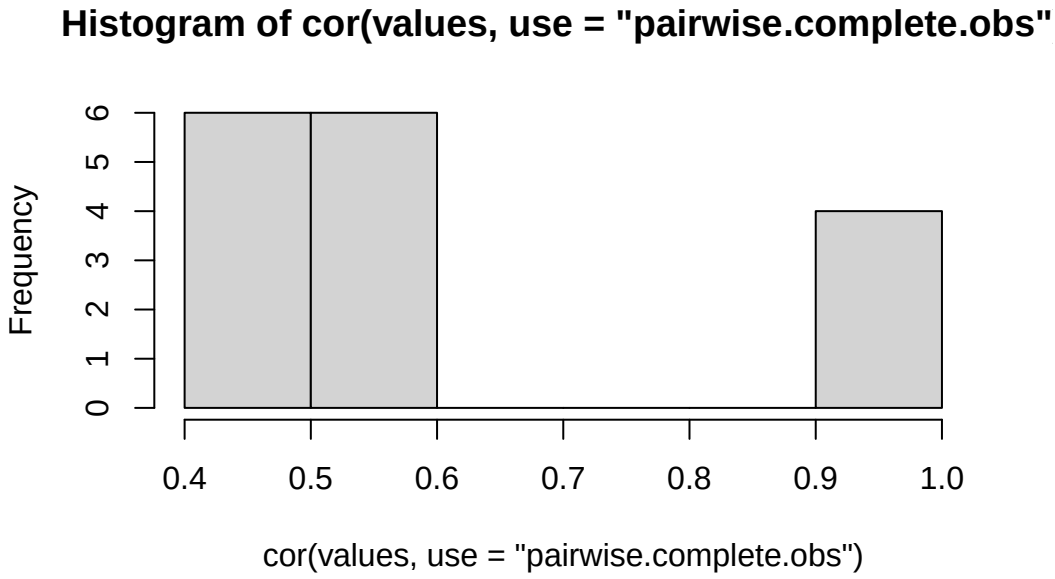


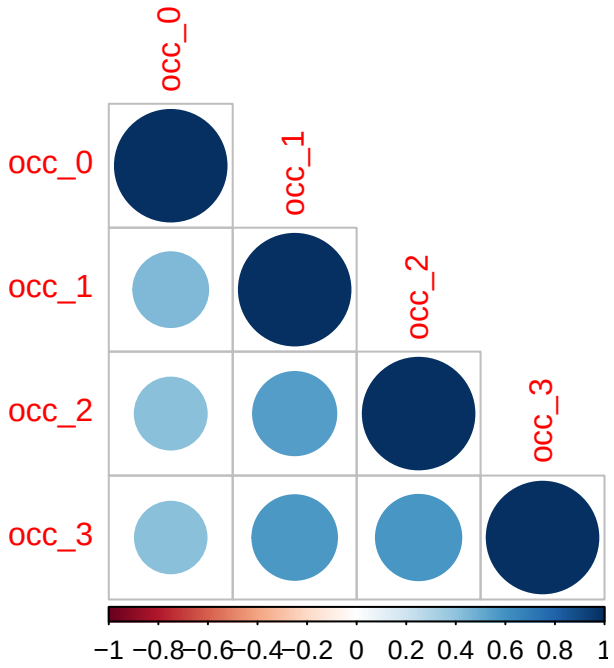


Secondary Outcome

This is raw correlation from the data. Looks like it it appropriate to estimate just one parameter for this matrix and go with exchangeable correlation structure. Average off-diagonal correlation is 0.4988736, which represents raw values adequately.

```
          occ_0     occ_1     occ_2     occ_3
occ_0 1.0000000 0.4495387 0.4155846 0.4114314
occ_1 0.4495387 1.0000000 0.5565082 0.5753455
occ_2 0.4155846 0.5565082 1.0000000 0.5848334
occ_3 0.4114314 0.5753455 0.5848334 1.0000000
```





4 - B

For sure, we need to use random intercepts. Might need to use random slopes, however, effectiveness of this modeling choice is not apparent here.

5

How will p-values and confidence intervals be computed? a. What will be used as the level of statistical significance? b. (Mixed models) Packages/options used in R/SAS for testing fixed and random effects. Note this can be included in your final report, but does not need to be included in your SAP as we wont have spent much time talking about this yet when you are writing your SAP.

5 -A

No multiple comparisons here, just use $\alpha = 0.05$.

will use geepack + emmeans + ggplot or

lme4 + lmeTest + emmeans + ggplot