## Oct 17, 2023 | [data play](https://www.google.com/calendar/event?eid=M25iaGhpcnBuNWFoZWJxMmNqdDhobWhlaG4gZ3VvZmFuZy5tYS5nckBkYXJ0bW91dGguZWR1)

Attendees: [Joseph Nano](mailto:joseph.nano.gr@dartmouth.edu) [J H](mailto:jane.han.gr@dartmouth.edu) [Freya Ma](mailto:guofang.ma.gr@dartmouth.edu)

**Meeting Agenda (30-45 minutes)**

* Data exploration - Excel
* Github Collaboration

**Copied from proposal**

**Aim 1:** **Visualization of insomnia symptomatology in adolescents by subgroups**

1.1. Visually identify distinct subgroups among adolescents based on demographic variables (e.g., **Age, Race**) and their distribution on the **GCTI, ASHS, ASQ** battery tests.

1.2. Utilize data visualization techniques (e.g., **heatmap analysis**) to depict the relationship between identified subgroups and psychological factors. Provide an intuitive representation of how diverse psychosocial factors contribute to varied sleep patterns in adolescents.

**Aim 2:** **Statistical Analysis of Sleep Quality and Habits in Adolescents**

2.1. Use **sleep quality** to predict **mental health and habits** among adolescents with regression analyses.

2.2. Investigate if the interaction effect of significant predictor factor(s) above and the race factor is statistically significant.

Comments from Ashlee

Your aims seem achievable and a great place to explore. I have only a few notes. 1) A little more detail would be helpful (ie, **why regression? What kind of regression model?** What is an outcome variable well reasoned from your research?).

Secondly, unless you have a really compelling reason to choose both R and Python; don't. Science is more reproducible when we use only the necessary tools. In some cases it's reasonable to mix languages because there isn't complete redundancy between R and Python so both are absolutely necessary. I doubt this will be the case for you, so I'd encourage you to focus on a very clear pipeline where each tool selected is well motivated by your research problem.

-> just stick to R

**Notes**

* Have missing data (NULL) :
* should we remove all null?
* What features do we want to focus on for our analysis?
* Should we make
* We will use ‘Sleep’ as a dependent variable
* Compare with other variables

**Completed task today**

* Cleaned the dataset (e.g., Replaced all ‘blanks’ & ‘\’ & ‘ ‘ to ‘NA’)
  + **Steps in Excel:** 
    - Edit → Find → Go to → Special → Blank → Type ‘NA’ and click ‘CTRL-return’ to replace all blanks into ‘NA’
  + **Next step:** we can check and get rid of all ‘NA’ in R and conduct further data wrangling
* Played around and generated with a few plots in Excel for reference

**Action items**

* **Future meetings:** 
  + Thursday, Oct.19th, 4pm-5pm (or 4:30pm)
  + Tuesday, Oct.24th, 4pm-5pm (or 4:30pm)
  + Thursday, Oct.26th, 4pm-5pm (or 4:30pm)
  + Tuesday, Oct.31st, 4pm-5pm (or 4:30pm)
  + Thursday, Nov.2nd, 4pm-5pm (Draft one presentation? & TA feedback)
  + Tuesday, Nov.7th, 4pm-5pm (\*\*Practice Presentation\*\*)
  + **Thursday, Nov.9th: FINAL PRESENTATION**