# Determining Data Wrangling Needs

# It all Starts with the Analysis!

- You need to know what analyses you’re using before you can determine data wrangling

- Check Data Wrangling and Visualization – Choosing the Right Statistic

- Understand the IVs and DVs for your questions
  - Categorical? How many levels?
  - Continuous?
  - How many of each?

# Look up the Assumptions and Requirements

- Review them in the LMS

- You can ask for assistance finding the particular location

# Some things to ask about your variables…

- Does the data need to be recoded?
- Does your data need to be a particular type?
- Do you need to categorize or re-categorize your data?
- Do you have time/date data that needs to be reformatted for usefulness?

# Some things to ask about your data…

- Do you have missing data to remove?
- Do you still have enough data once you remove rows with missing information?
- Do you have too much data, and need to remove some to increase processing speeds?
- Do any columns need to be renamed? Are there spaces in your columns?
- Is the data in the right shape for the analysis, or will it need to be transformed?

# Do you have multiple datasets?

- Do you want to add columns (merge) or rows (append)?
- Is there a common link between them?
- Is that link unique? Or will you need to aggregate data?
- Do they have all the same columns?
- Are the columns all named the same?

# Assumptions to think about

- Normality
- Multicollinearity
- Sample size
- Homoscedasticity
- Homogeneity of variance
- Outliers
- Linearity

# If you want to do visuals in Python / R…

- Does your data need to be in a particular format for the visual?

- Do you require a particular data type?

# Stay Organized

- Create cards in Trello for each task
- Have separate cards for “determine what to do” and “do it”
- Add in cards for study time or refreshing

# Questions?