**Apply data wrangling to your capstone project**

**NCAA Financials**

This was a relatively straight forward dataset. Each row is a year of NCAA financial information for a specific NCAA participating university. In wrangling this dataset, I removed any text columns aside from the university name. The unique identifier provided in this data set is the column “united” which is assigned to each university. I grouped the data set by the variable “chronname” after converting some columns being stored as factors though they contained integer data to get averages for the most meaningful columns. I then found the averages for the most meaningful/useable variables in the data set including subsidy amount, student tuition, Athletic revenues and fees. Original dataset had over 40 unique variables

**Socio Economic Mobility Data**

This study utilized 11 different data sets with 1 dataset as an index. This was a particularly challenging task because not all datasets were pertaining to the information I was looking for regarding socio economic mobility rates as clearly as I would have like. Of the 10 datasets, 4 provide really useful information in reference to this project and were relatively.

*Table 3: Baseline Longitudinal Estimates by College and Child’s Cohort*

Socio economic motility rates are recorded and summarized by original cohort class (1980 – 1991). Blank values in this data frame reflect that a class did not exist (blank is also reflected in “count” column. Count references the number of people in a particular cohort attending a particular university). Replaced blanks with “NA”.

*Table 4: Cross-Sectional Estimates by College: Heterogeneity by Gender and Alternative College and Income Definitions*

This is the most robust data set of the 4 and used a different set of alternative college and income definitions. Left this dataset as is. Data is also organized by child gender, and adjusts for regional cost of living differences when measuring incomes.

*Table 6: Cross-Sectional Statistics by College Tier and Parent Income Percentile*

Child income information is presented at each parental socioeconomic percentile across 15 university tiers identified by study. Dataset can provide more general trends across the difference income percentiles/ranks. Created a DF that has the averages all values across columns grouped by parent’s income percentile, regardless of what tier college the student attended. This may provide some general overarching stats about what a child’s mobility rates are just based on what income percentiles their parents are in.

*Table 10: College Level Characteristics*

Race, majors, SAT, graduation rates, baron’s index of selectivity, graduation and college tuition costs are all captured in this data set. All data is organized by superID and will be needed for Shiny Application of project. Left dataframe as is.