initial exploration

ccid <- read\_csv(here::here("data/cc\_institution\_details.csv"))

## Warning: 34 parsing failures.  
## row col expected actual file  
## 1041 ft\_fac\_value a double NULL '/Users/jasongrahn/da485\_capstone/data/cc\_institution\_details.csv'  
## 1041 ft\_fac\_percentile a double NULL '/Users/jasongrahn/da485\_capstone/data/cc\_institution\_details.csv'  
## 1055 ft\_fac\_value a double NULL '/Users/jasongrahn/da485\_capstone/data/cc\_institution\_details.csv'  
## 1055 ft\_fac\_percentile a double NULL '/Users/jasongrahn/da485\_capstone/data/cc\_institution\_details.csv'  
## 1504 ft\_fac\_value a double NULL '/Users/jasongrahn/da485\_capstone/data/cc\_institution\_details.csv'  
## .... ................. ........ ...... ..................................................................  
## See problems(...) for more details.

ccid\_var <- ccid %>%  
 #removing unneeded variables  
 select(-(contains("percentile"))) %>%   
 select(-(starts\_with("vsa\_"))) %>%   
 select(-site, -similar, -nicknames, -counted\_pct, -cohort\_size, -hbcu,   
 -awards\_per\_state\_value, -awards\_per\_natl\_value, -exp\_award\_state\_value,   
 -exp\_award\_natl\_value, -grad\_100\_value, -grad\_150\_value, -state\_sector\_ct,   
 -carnegie\_ct) %>%   
 #convert flagship to binary.   
 mutate(flagship\_binary = case\_when(flagship == "X" ~ 1,  
 flagship == "NULL" ~ 0),  
 #convert percentages to decimal values  
 awards\_per\_value = awards\_per\_value/100,  
 ft\_pct = ft\_pct / 100,  
 pell\_value = pell\_value / 100,  
 ft\_fac\_value = ft\_fac\_value / 100,  
 retain\_value = as.numeric(retain\_value) / 100)

## Warning: NAs introduced by coercion

write\_csv(x = ccid\_var,  
 path = "ccid\_var.csv")

## Cleanup notes:

* percetiles should be converted to decimals? making a list of these.
* binaries from X and NULL to 1 and 0

## The data includes the following variables. I’m re-grouping them by the ones I think may provide value for the niche area we’re studing.

* unitid - Department of Education Unit ID number
* chronname - Institution name
* city - Institution city
* state - Institution state
* level - Level of institution (4-year, 2-year) [[We might think of only keeping 4 year?]]
* control - Control of institution (Public, Private not-for-profit, Private for-profit)
* basic - Carnegie Foundation for the Advancement of Teaching Basic Classification (2010 version)
* flagship - Denotes Public flagship institutions
* flagship\_binary - This is the flagship variable converted to 1 and 0.
* long\_x - Institution longitude
* lat\_y - Institution latitude
* student\_count - Total number of undergraduates in 2010 [[maybe a measure of GIGO?]]
* awards\_per\_value - Completions per 100 FTE undergraduate students (average 2011, 2012, and 2013) [[needs to convert to decimal]]
* exp\_award\_value - Estimated educational spending (expenses related to instruction, research, public service, student services, academic support, institutional support, operations and maintenance) per academic award in 2013. Includes all certificates and degrees
* ft\_pct - Percentage of undergraduates who attend full-time [[needs to convert to decimal]]
* fte\_value - total number of full-time equivalent undergraduates
* med\_sat\_value - Median estimated SAT value for incoming students
* aid\_value - The average amount of student aid going to undergraduate recipients
* endow\_value - End-of-year endowment value per full-time equivalent student
* pell\_value - percentage of undergraduates receiving a Pell Grant [[needs to convert to decimal]]
* retain\_value - share of freshman students retained for a second year [[needs to convert to decimal]]
* ft\_fac\_value - Percentage of employees devoted to instruction, research or public service who are full-time and do not work for an associated medical school [[needs to convert to decimal]]

## Removed:

* hbcu - Denotes Historically Black Universities
* awards\_per\_state\_value - Completions per 100 FTE undergraduate students, state and sector average maybe unnecessary - compares colleges, not students.
* awards\_per\_natl\_value - Completions per 100 FTE undergraduate students, national sector average maybe unnecessary - compares colleges, not students.
* exp\_award\_state\_value - Spending per completion, state average maybe unnecessary - compares colleges, not students.
* exp\_award\_natl\_value - Spending per completion, national average maybe unnecessary - compares colleges, not students.
* med\_sat\_percentile - Institution’s percent rank for median SAT value within sector maybe unnecessary - compares colleges, not students.
* aid\_percentile - Institution’s percent rank for average amount of student aid going to undergraduate recipients within sector maybe unnecessary
* endow\_percentile - Institution’s percent rank for endowment value per full-time equivalent student within sector maybe unnecessary
* grad\_100\_value - percentage of first-time, full-time, degree-seeking undergraduates who complete a degree or certificate program within 100 percent of expected time (bachelor’s-seeking group at 4-year institutions)
* grad\_100\_percentile - Institution’s percent rank for completers within 100 percent of normal time within sector
* grad\_150\_value - percentage of first-time, full-time, degree-seeking undergraduates who complete a degree or certificate program within 150 percent of expected time (bachelor’s-seeking group at 4-year institutions)
* grad\_150\_percentile - Institution’s percent rank for completers within 150 percent of normal time within sector
* pell\_percentile - Institution’s percent rank for undergraduate Pell recipients within sector maybe unnecessary
* retain\_percentile - Institution’s percent rank for freshman retention percentage within sector
* ft\_fac\_percentile - Institution’s percent rank for full-time faculty share within sector maybe unnecessary
* vsa\_year - Most recent year of Student Success and Progress Rate data available
* vsa\_grad\_after4\_first - First-time, full-time students who graduated from this institution within four years
* vsa\_grad\_elsewhere\_after4\_first - First-time, full-time students who graduated from another institution within four years — Did they LEAVE to graduate? Maybe unnecessary, but good context.
* vsa\_enroll\_after4\_first - First-time, full-time students who are still enrolled at this institution after four years maybe unnecessary
* vsa\_enroll\_elsewhere\_after4\_first - First-time, full-time students who are enrolled at another institution after four years
* vsa\_grad\_after6\_first - First-time, full-time students who graduated from this institution within six years Maybe unnecessary, but good context.
* vsa\_grad\_elsewhere\_after6\_first - First-time, full-time students who graduated from another institution within six years Maybe unnecessary, but good context.
* vsa\_enroll\_after6\_first - First-time, full-time students who are still enrolled at this institution after six years Maybe unnecessary, but good context.
* vsa\_enroll\_elsewhere\_after6\_first - First-time, full-time students who are enrolled at another institution after six years Maybe unnecessary, but good context.
* vsa\_grad\_after4\_transfer - Full-time transfer students who graduated from this institution within four years maybe unecessary
* vsa\_grad\_elsewhere\_after4\_transfer - Full-time transfer students who graduated from another institution within four years maybe unecessary
* vsa\_enroll\_after4\_transfer - Full-time transfer students who are still enrolled at this institution after four years maybe unecessary
* vsa\_enroll\_elsewhere\_after4\_transfer - Full-time transfer students who are enrolled at another institution after four years maybe unecessary
* vsa\_grad\_after6\_transfer - Full-time transfer students who graduated from this institution within six years
* vsa\_grad\_elsewhere\_after6\_transfer - Full-time transfer students who graduated from another institution within six years
* vsa\_enroll\_after6\_transfer - Full-time transfer students who are still enrolled at this institution after six years
* vsa\_enroll\_elsewhere\_after6\_transfer - Full-time transfer students who are enrolled at another institution after six years
* counted\_pct - share entering undergraduate class who were first-time, full-time, degree-seeking students, meaning that they generally would be part of a tracked cohort of potential graduates. The entering class of 2007 is displayed for 4-year institutions; 2010 for 2-year institutions.