

# Group Project I - College Scorecard Dataset

*Kevin Hunt, Noah Johnson, Lydia Laseur, Michael McCormack*

*November 4, 2017*

## College Scorecard Data Aggregation and Cleaning

To view the R code for cleaning the College Scorecard Dataset, assessing the quality of the data, and adding latitude, longitude and rurality see the `clean_all.R` file.

The we have chosen 136 columns to work with:

---

year	program_percentage.parks_recreation_fitness
id	program_percentage.philosophy_religious
ope8_id	program_percentage.theology_religious_vocation
ope6_id	program_percentage.physical_science
name	program_percentage.science_technology
city	program_percentage.psychology
state	program_percentage.security_law_enforcement
zip	program_percentage.public_administration_social_service
location.lat	program_percentage.social_science
location.lon	program_percentage.construction
degrees_awarded.predominant_recoded	program_percentage.mechanic_repair_technology
degrees_awarded.predominant	program_percentage.precision_production
degrees_awarded.highest	program_percentage.transportation
locale	program_percentage.visual_performing
program_percentage.agriculture	program_percentage.health
program_percentage.resources	program_percentage.business_marketing
program_percentage.architecture	program_percentage.history
program_percentage.ethnic_cultural_gender	attendance.academic_year
program_percentage.communication	tuition.in_state
program_percentage.communications_technology	tuition.out_of_state
program_percentage.computer	faculty_salary
program_percentage.personal_culinary	share_25_older
program_percentage.education	title_iv.died_by.2yrs
program_percentage.engineering	title_iv.low_inc.died_by.2yrs
program_percentage.engineering_technology	title_iv.mid_inc.died_by.2yrs
program_percentage.language	title_iv.high_inc.died_by.2yrs
program_percentage.family_consumer_science	title_iv.depend.died_by.2yrs
program_percentage.legal	title_iv.independ.died_by.2yrs
program_percentage.english	title_iv.female.died_by.2yrs
program_percentage.humanities	title_iv.male.died_by.2yrs
program_percentage.library	title_iv.pell_recip.died_by.2yrs
program_percentage.biological	title_iv.no_pell.died_by.2yrs
program_percentage.mathematics	title_iv.loan_recip.died_by.2yrs
program_percentage.military	title_iv.no_loan.died_by.2yrs
program_percentage.multidiscipline	title_iv.first_gen.died_by.2yrs

---

title_iv.not_first_gen.died_by.2yrs	title_iv.depend.died_by.6yrs
title_iv.died_by.3yrs	title_iv.independ.died_by.6yrs
title_iv.low_inc.died_by.3yrs	title_iv.female.died_by.6yrs
title_iv.mid_inc.died_by.3yrs	title_iv.male.died_by.6yrs
title_iv.high_inc.died_by.3yrs	title_iv.pell_recip.died_by.6yrs
title_iv.depend.died_by.3yrs	title_iv.no_pell.died_by.6yrs
title_iv.independ.died_by.3yrs	title_iv.loan_recip.died_by.6yrs
title_iv.female.died_by.3yrs	title_iv.no_loan.died_by.6yrs
title_iv.male.died_by.3yrs	title_iv.first_gen.died_by.6yrs
title_iv.pell_recip.died_by.3yrs	title_iv.not_first_gen.died_by.6yrs
title_iv.no_pell.died_by.3yrs	title_iv.died_by.8yrs
title_iv.loan_recip.died_by.3yrs	title_iv.still_enrolled_by.8yrs
title_iv.no_loan.died_by.3yrs	title_iv.low_inc.died_by.8yrs
title_iv.first_gen.died_by.3yrs	title_iv.mid_inc.died_by.8yrs
title_iv.not_first_gen.died_by.3yrs	title_iv.high_inc.died_by.8yrs
title_iv.died_by.4yrs	title_iv.depend.died_by.8yrs
title_iv.low_inc.died_by.4yrs	title_iv.independ.died_by.8yrs
title_iv.mid_inc.died_by.4yrs	title_iv.female.died_by.8yrs
title_iv.high_inc.died_by.4yrs	title_iv.male.died_by.8yrs
title_iv.depend.died_by.4yrs	title_iv.pell_recip.died_by.8yrs
title_iv.independ.died_by.4yrs	title_iv.no_pell.died_by.8yrs
title_iv.female.died_by.4yrs	title_iv.loan_recip.died_by.8yrs
title_iv.male.died_by.4yrs	title_iv.no_loan.died_by.8yrs
title_iv.pell_recip.died_by.4yrs	title_iv.first_gen.died_by.8yrs
title_iv.no_pell.died_by.4yrs	title_iv.not_first_gen.died_by.8yrs
title_iv.loan_recip.died_by.4yrs	share_firstgeneration
title_iv.no_loan.died_by.4yrs	share_firstgeneration_parents.middleschool
title_iv.first_gen.died_by.4yrs	share_firstgeneration_parents.highschool
title_iv.not_first_gen.died_by.4yrs	share_firstgeneration_parents.somecollege
title_iv.died_by.6yrs	demographics.age_entry
title_iv.low_inc.died_by.6yrs	demographics.over_23_at_entry
title_iv.mid_inc.died_by.6yrs	demographics.first_generation
title_iv.high_inc.died_by.6yrs	rurality

When analyzing the data, please keep in mind that not every year has good completeness of the data. Use `completeness.by.year` to explore the data before you begin visualization. Example:

```
completeness.by.year[, c("year", "locale")]
```

year	locale
1996_97	0.0000000
1997_98	0.0000000
1998_99	0.0000000
1999_00	0.0000000
2000_01	0.0000000
2001_02	0.0000000
2002_03	0.0000000
2003_04	0.0000000
2004_05	0.0000000
2005_06	0.0000000

year	locale
2006_07	0.0000000
2007_08	0.0000000
2008_09	0.0000000
2009_10	0.0000000
2010_11	0.0000000
2011_12	0.0000000
2012_13	0.0000000
2013_14	0.0000000
2014_15	0.0000000
2015_16	0.9412617

## Visualizing the College Scorecard Dataset

Please select a question (or come up with your own) and add some visualizations to display any relationships of interest. Most questions we came up with involved the rurality of the schools. The original dataset includes this information for the 2015-2016 school year in the “locale” column. For all the years, I’ve added a column of rurality to the dataset (original data can be found in ziprural.txt or downloaded [here](#)). The dictionary for explanations of these codes is as follows:

A1	Rural Urban Continuum Code		
Code	Description	Counties	2000 Population
<b>Metro Counties:</b>			
1	Counties in metro areas of 1 million population or more	413	149,224,067
2	Counties in metro areas of 250,000 to 1 million population	325	55,514,159
3	Counties in metro areas of fewer than 250,000 population	351	27,841,714
<b>Nonmetro Counties:</b>			
4	Urban population of 20,000 or more, adjacent to a metro area	218	14,442,161
5	Urban population of 20,000 or more, not adjacent to a metro area	105	5,573,273
6	Urban population of 2,500 to 19,999, adjacent to a metro area	609	15,134,357
7	Urban population of 2,500 to 19,999, not adjacent to a metro area	450	8,463,700
8	Completely rural or < 2,500 urban population, adjacent to a metro area	235	2,425,743
9	Completely rural or < 2,500 urban population, not adjacent to a metro area	435	2,802,732

Questions we discussed in class:

**Rurality story:** Where did they come from? Where did they go? Where did they come from, **Cotton Eye Joe**? Do more rural schools show a higher percentage of returning adults than metro schools? Do more rural schools have a higher percentage of first generation students? How do the degrees awarded differ between rurality classes (do rural areas award more technical degrees)? Are more rural schools cheaper?

**Other Questions** How does the average faculty salary of a school scale with the cost of living in that area? (requires finding cost of living)  
Dangerous schools/regions with the death statistics