# Proposed usage of datasets

Logistic regression unit

* Lecture/lab: principal.csv
* Assignment: nscs.csv / portugal.csv

Poisson regression unit

* Lecture/lab: ceps\_sch.csv
* Assignment: ah\_alcohol.csv (add health dataset in progress; outcome variable is a measure of days of alcohol consumption over the past year) / portugal.csv

Nested data unit

* Lecture/lab: seda.csv
* Assignment: gpa.csv / gpa\_above.csv

measurement

* Lecture/lab: principal.csv
* Assignment: ah\_happiness.csv (add health dataset in progress; a set of survey questions about happiness)

Missing data

* Lecture/lab: galo.csv
* Assignment: datasets from previous sessions

Final project

* Add Health datasets (add health datasets in progress)
* MOCCA sample dataset from Gina

# Newly developed datasets

## China Education Panel Survey (ceps\_sch.csv)

This dataset was drawn from the [China Education Panel Survey (CEPS)](http://ceps.ruc.edu.cn/English/Overview/Overview.htm), China’s first nationally representative, longitudinal survey of middle school students. CEPS first wave was conducted during school year 2013-2014 and a stratified, four-step random sampling procedure was implemented to draw a random sample of middle schools (N=112) from the nation and collect rich data from school administrators, teachers, students, and parents. This school-level dataset is limited to 45 schools that have eight teacher participants and no missing values on any key variables. Our primary goal with this dataset is to investigate the school-level factors associated with teacher job satisfaction that is measured by the count of teachers who feel satisfied with their job and their school.

* schids, school identifier
* njobsatis, number of teachers who answered yes to the question of whether to choose being a teacher if given other job options
* nschsatis, number of teachers who answered yes to the question of whether to choose the current school if being allowed to move to other schools
* rural, categorical measure of proportion of rural students in school; 1=lower than 25%, 2=25% to 60%, 3=60% to 80%, 4=higher than 80%
* clsize, average class size
* computer, number of computers available to students
* ctyeduz, neighborhood average years of education (standardized)

## Portuguese student survey (portugal.csv)

Statistics keep Portugal at Europe’s tail end due to its high student failure rates especially in core content areas such as math and the Portuguese language. Cortez and Silva (2008) collected student data from two Portuguese schools during the 2005- 2006 school year using school reports and student survey with a goal of developing models to predict student performance. The data include student grades, demographic, social, and school related features. Our goal in this class is to investigate the factors associated with student failure in Portuguese and the magnitudes of these associations. Note that student Portuguese grade is measured on a 20-point scale with zero being the lowest and 20 being the perfect grade. The grade can be further measured by a five-level classification system, with 16-10 being excellent, 14-15 being good, 12-13 being satisfactory, 10-11 being sufficient, and 0-9 being fail.

Potentially, since the association between two variables are bidirectional, our investigations can explore another direction. For example, what is the association with student failing Portuguese/grade in Portuguese in first period and tutoring/motivation/alcohol consumption/quality of family relationships by the time of survey?

The dataset contains 649 student observations and 18 variables, detailed below:

* school, student’s school (‘GP’ - Gabriel Pereira; ‘MS’ - Mousinho da Silveira)
* female, binary variable coded one for female students
* age, student’s age
* cohab, binary variable coded one if parents are living together with the student
* mjob, mother’s job (‘teacher’, ‘health’ care related, civil ‘services’ (e.g. administrative or police), ‘at\_home’, or ‘other’)
* fjob, father’s job (‘teacher’, ‘health’ care related, civil ‘services’ (e.g. administrative or police), ‘at\_home’, or ‘other’)
* tutoring, binary variable coded one if the student is having afterschool, paid tutoring on Portuguese
* motivation, binary variable coded one if the student wants to take higher education
* romantic, binary variable coded one if the student is in a romantic relationship
* spr, five-point likert scale measure of the quality of family relationships (from 1=very bad to 5=excellent)
* health, five-point likert scale measure of the student’s current health status (from 1=very bad to 5=very good)
* alc, five-point likert scale measure of the student’s weekly alcohol consumption (from 1=very low to 5=very high)
* por1, first period grade
* por2, second period grade
* por3, final grade
* por1\_fail, binary variable coded one for students who failed Portuguese in first period
* por2\_fail, binary variable coded one for students who failed Portuguese in second period
* por3\_fail, binary variable coded one for students who failed Portuguese at the end of school year

## Add Health large dataset for final project and potentially sub-datasets for labs (addhealth.csv)

The dataset is individual level data of 3,418 Add Health study participants who participated in wave 1, wave 3, and wave 4 surveys. It has 34 variables besides the identifier variable.

**Predictor and control variable candidates from Wave 1**:

* id, the individual’s identifier
* female, dummy variable coded one for individuals whose biological sex is female
* hispanic, white, black, native, and asian are all dummy variables coded one for individuals whose self-identified race is Hispanic or Latino, White, Black or African American, American Indian or Native American, and Asian or Pacific Islander, respectively
* language, dummy variable coded one for individuals whose home language is English
* grade, factor variable indicating the individual’s grade (7, 8, 9, 10, 11, or 12)
* medu, fedu, and prtedu represent the highest years of education of the individual’s household mother, father, and both parents, respectively. The coding procedure is attached at the end.
* health, the individual’s self-report health status; 1=excellent, 2=very good, 3=good, 4=fair, 5=poor, 6=refused, 8=don’t know
* sleepenough, the individual’s self-report rating of whether they get enough sleep; 0=never, 1=just a few times, 2=about once a week, 3=almost every day, 4=every day
* repeating, dummy variable coded one for individuals who ever repeated a grade
* skipping, dummy variable coded one for individuals who ever skipped a grade
* oos, dummy variable coded one for individuals who got at least one out-of-school suspension
* expelled, dummy variable coded one for individuals who were ever expelled
* tsr and ssr measure the individual’s self-rating on the questions “Are you getting along with your teachers?” and “Are you getting along with other students?”; 0=never, 1=just a few times, 2=about once a week, 3=almost everyday, 4=everyday, 6=refused, 7=ligitimate skip, 8=don’t know

**Outcome variables from waves 3 and 4**

* education, the individual’s total years of education
* gpa, the individual’s high school final GPA points
* gpa\_3cat, the individual’s high school final GPA points coded into three categories: 0=below 2.0, 1=above 2.0 but less than 3.0, 2=above 3.0
* gpa\_5cat, the individual’s high school final GPA points coded into five categories: 0=below 2.0, 1=above 2.0 but less than 3.0, 2=above 3.0 but less than 3.4, 3=above 3.4 but less than 3.8, 4=above 3.8
* income, the individual’s annual earnings (including salary and all types of other incomes)
* income\_log, the individual’s log-transformed annual earnings
* jail, dummy variable coded one for individuals who had ever been in a jail, prison, juvenile detention center or other correctional facility
* suicidethought, dummy variable coded one for individuals who had ever seriously thought about committing suicide during the past 12 months by the time of survey
* q1 to q8, eight survey questions measured on a four-point likert-scale, detailed below:
  + Question descriptions: “Think about the past seven days. How often was each of the following things true during the past seven days?”
    - q1, “You were bothered by things that usually don’t bother you.”
    - q2, “How often do you feel isolated from others?”
    - q3, “You had trouble keeping your mind on what you were doing.”
    - q4, “You felt depressed.”
    - q5, “You felt happy.”
    - q6, “You enjoyed life.”
    - q7, “You felt sad.”
    - q8, “You felt that people disliked you.”
  + Levels: 0 = Never, 1 = Sometimes, 2 = A lot of the time 3 = Most of the time or all of the time 6 = Refused 8 = Don’t know

For years of education variables, follow Kraft et al working paper p.14, the coding is as following:

* “8th grade or less” = 8 years
* “some high school” = 11 years
* “high school graduate” = 13 years
* “some vocational/technical training (after high school) = 13.5 years
* “completed vocational/technical training (after high school)” = 14
* “some college” = 15
* “completed college (bachelor’s degree)” = 17 years
* “some graduate school” and “some post baccalaureate professional education = 18 years
* “completed a master’s degree” = 19 years
* “some graduate training beyond a master’s degree” and “completed post baccalaureate professional education = 20 years
* “completed a doctoral degree” = 22 years

## National school counselor survey (nscs.csv)

To gain a broad picture of the counselor experience during the early months of COVID-19 school closures, [Savitz-Romer and Rowan-Kenyon (2021)](https://doi.org/10.3886/E145681V1-98740) created an 80-question online survey to collect data from 1,060 school counselors and educators in adjacent roles (e.g., college counselors, adjustment counselors, counseling directors) in 48 states and Puerto Rico in the US. For simplicity reasons, we only include in our dataset the key variables as following (dropping incomplete observations on these variables will result in N = 787):

* id, counselor identifier
* grades, school level, 0 = elementary, 1 = middle school, 2 = high school, 3 = mixed
* schooltype, school type, 1 = public, 2 = charter, 3 = private
* urbanicity, school urbanicity level, 1 = urban, 2 = suburban, 3 = rural
* female, counselor gender coded one for female and zero for others
* exp, counselor experience measured by years
* ooo, dichotomous variable coded one if the counselor had less time for one-on-one counseling due to COVID
* group, dichotomous variable coded one if the counselor had less time for group counseling due to COVID
* sel, dichotomous variable coded one if the counselor had less time for supporting student social emotional needs due to COVID

## School principal data (principal.csv)

This dataset was drawn from Meta Krüger’s (1994) study investigating gender differences in school leadership in Netherlands. The study implemented a matching procedure to generate pairs of comparable schools (N=98) with the only difference being that one school principal was female and the other male, then surveyed the school principals, teachers, and students. This student-level dataset contains gender information on school principal and nine randomly selected students from each school, and measures on student-perceived relationship with their principal.

The dataset contains 8 variables, detailed below.

* schid, school identifier
* stuid, student identifier within each school
* pfemale, binary variable coded one for female principals
* page (fake variable for teaching purposes), principal’s age
* sfemale, binary variable coded one for female students
* match, coded one if the student and their principal are the same gender
* attention: take the average score of student’s agreements (rated on a 4-point likert scale from 1=low to 4=high) with questions “Sometimes the principal talks to me”, “I think the principal knows who I am”, and “Principal knows how well I am doing”, then standardize it at school level to proxy for student-perceived attention from principal
* affective: a binary variable coded one for students who rated somewhat high or high on “the school principal is nice” to proxy for student’s attitude toward principal

## College GPA data (gpa.csv and gpa\_above.csv)

The GPA data is a longitudinal dataset, where 200 college students were followed for six consecutive semesters. The data was simulated. Note that this dataset contains incomplete observations.

For gpa.csv, key variables are:

* stuid, student identification number
* time, a factor variable indexing the six consecutive semesters
* gpa, student’s GPA at the end of semester
* female, coded one for female students and zero for male
* hsgpa, high school GPA

For gpa\_above.csv, key variables are:

* stuid, student identification number
* gpa\_above, a binary variable coded one for students whose overall college GPA was above 3.0 and zero otherwise
* female, coded one for female students and zero for male
* hsgpa, high school GPA

## GALO data (particularly for Unit 5 dealing with missing data) (galo.csv)

The GALO data is originally described by Peschar (1975) and analyzed by Dronkers and Schijf (1994) among other studies. The 1959 cohort consists of 1270 school children in the sixth grade of 37 elementary schools in the city of Groningen (Netherlands).

Variables are detailed below.

* schid, school identification number
* female, coded one for female students and zero for male
* galo, student’s achievement score on GALO test
* advice, teacher’s advice about secondary education; 0=“no subsequent school”, 1=“lowest”,…, 6=“highest”, 999=“missing”
* medu, mother’s highest education; 1=“lowest”,…, 9=“highest”, 999=“missing”
* fedu, father’s highest education; 1=“lowest”,…, 9=“highest”, 999=“missing”
* focc, father’s occupational status; 1=“lowest”,…, 6=“highest”, 9=“missing”

## Binge Eating in Men

This dataset is the original dataset used in a study conducted by Dr. Nichole Kelly and her colleague ([Kelly, Cotter, & Guidinger, 2018](https://pubmed.ncbi.nlm.nih.gov/29990652/)). They examined whether perceived overeating patterns were uniquely associated with eating- and weight-related comorbidities and found that young men who engage in both subjective (SBEs) and objective binge eating episodes (OBEs) may be at the highest risk for chronic disease and psychological concerns.

Analytic Sample. This dataset is individual-level data for 1114 young men (age 18-30), who were recruited from across US and finished the survey through a link sent via email. The data contains 343 incomplete observations.

The data set contains 13 variables, detailed below.

* *none*, dummy variable coded 1 for individuals without disordered eating
* *oe*, dummy variable coded 1 for individuals with OE(s) only; note: OE means overeating without presence of loss of control
* *sbe*, dummy variable coded 1 for individuals with SBE(s) only; note: SBE means subjective binge eating
* *obe*, dummy variable coded 1 for individuals with OBE(s) only; note: OBE means objective binge eating
* *sobe*, dummy variable coded 1 for individuals with both OBE(s) and SBE (s)
* *dietary*, mean dietary restraint score (higher scores = higher restraint)
* *emotion*, measure of emotion dysregulation problems (higher scores = more problems)
* *image*, body image concern measure, higher = greater internalization of thin and muscular ideal
* *race*, Racial/ethnic identity (four categories)
* *bmi, Body mass index (in kg/m-squared)*
* *age*
* *edu, education group (nine categories)*
* *employ*, employment status (four categories)

## CEPS

I have access to and knowledge about [China Education Panel Survey (CEPS)](http://ceps.ruc.edu.cn/English/Overview/Overview.htm), China’s nationally representative, longitudinal survey of middle school students. Two-year data on 9,449 students from the initial 7th grade cohort is publicly available. CEPS data contains 400+ school-level, 200+ teacher-level, 300+ parent-level, and 400+ student-level variables. We could talk more about what we need for this class, including lectures and assignments, and I can work on the datasets accordingly.

# Datasets used across EDUC 641 & 643

## 1. Teacher Professional Development Study (PDS)

This dataset is drawn from the NCRECE [Teacher Professional Development Study (PDS)](https://www.icpsr.umich.edu/web/ICPSR/studies/34848) database. This study was a randomized controlled evaluation of two forms of professional development (PD) - coursework (phase 1) and consultancy (phase 2) - delivered to about 490 early childhood education teachers across the nation. These PD supports aimed to improve teachers’ implementation of language/literacy activities and interactions with children, as well as promote gains in children’s social and academic development.

Our dataset is student-level data for 440 preschool students. Since PDS student outcome data has not been collected during phase 1, we focus on phase 2 intervention. Specifically, the sample students were in phase 1 treatment group and in either the treatment or control group during phase 2. Observations with missing values on any of the key variables were deleted for simplification reasons.

The data set contains 11 variables, detailed below.

* *childid*, unique identification number for the student
* *tchid*, unique identification number for the student’s teacher
* *schid*, unique identification number for the student’s school
* *cohort*, coded 1 for cohort 1 students (who started the experiment in spring 08) and 2 for cohort 2 students (who started the experiment in spring 09)
* *treat*, coded 1 for students in phase 2 treatment group and 0 for those in phase 2 control group
* *vocabulary*, the student’s post phase 2 score on a receptive vocabulary test (Peabody Picture Vocabulary Test-3rd edition)
* *female*, coded 1 for female students and 0 for male students
* *language\_eng*, coded 1 for students whose primary language is English and 0 otherwise
* *disability*, coded 1 for students with a disability and 0 otherwise
* *ethnicity*, eight-category variable documenting the student’s ethnicity
* *mother\_edu*, the student’s mother’s total years of education

## 2. Add Heath

This dataset is drawn from the third and fourth waves of National Longitudinal Study of Adolescent to Adult Health ([Add Health](https://addhealth.cpc.unc.edu/data/#public-use)) public-use data. Add Health is a longitudinal study of a nationally representative sample of over 20,000 adolescents (the public-use data sample is much smaller in size) who were in grades 7-12 during the 1994-95 school year, containing rich demographic, social, familial, socioeconomic, behavioral, psychosocial, cognitive, and health information. When the third (2001) and fourth (2008) waves took place, the participants were aged from 18-26 and from 24-32.

[Kraft et al (2021)](https://www.edworkingpapers.com/ai21-441) took advantage of this data and examined an understudied area - the pathway through which mentoring relationship between students and school personnel influences human capital development. Particularly, they found that having a school-based mentor is associated with an increase of 0.24 GPA points at high-school, a full year of additional education, and an increase of $60,600-$92,400 in lifetime earnings.

The dataset you’ll be using is individual-level data for Add Health participants who have no missing data on our variables of interested in both waves 3 and 4. Note that for simplifying reasons, we use smaller datasets than Kraft et al does and the measures of predictor and outcome variables are slightly different. Also we do not ask you to fit the fixed-effects models in Kraft et al, therefore all the results you get from this exercise cannot be interpreted as causal effects.

**Key variables:**

* id*, the individual’s unique identification number.*
* mentor*, binary variable, coded 1 for individuals who reported to have a school-based mentor (teachers/guidance counselors and coaches/athletic directors). This variable exists only in ah01.csv dataset.*
* mentee\_age*, the individual’s age when the mentor started to have an impact in their life. This variable exists only in ah02.csv dataset.*
* education*, the individual’s total years of education (recoded from Add Health data using the same approach in Kraft et al p.14).*
* income\_log*, the individual’s log-transformed annual earnings (including salary and all types of other incomes).*
* gpa*, the individual’s high school final GPA points.*
* gpa\_3*, binary variable coded 1 for individuals whose high school final GPA was above 3.0*

## 3. Stanford Education Data Archive (SEDA)

The data set we’ll be using in our final project was drawn from the Stanford Education Data Archive ([SEDA](https://edopportunity.org)) version 4.1. SEDA was launched in 2016 to provide nationally comparable, publicly available test score data for U.S. public school districts, allowing scientific inquiries on the relationships between educational conditions, contexts, and outcomes (especially student math/ELA achievements) at the district-level across the nation. It contains rich variables including measures of academic achievement and achievement gaps for school districts and counties, as well as district-level measures of racial and socioeconomic composition, racial and socioeconomic segregation patterns, and other features of the schooling system. Some descriptive findings can be found [here](https://edopportunity.org/discoveries/).

Due to the large size of SEDA full data set, we focus on the school year 2017-18 data for the state of Oregon. Specifically, our data set is district-level data for 103 Oregon school districts. Observations with missing values on any of the key variables were deleted for simplification reasons.

The data set contains 13 variables, detailed below.

* *district*, name of the district
* *subject*, coded “mth” fo Mathematics and “rla” for English/Language Arts
* *grade*, coded 3, 4, 5, 6 for grades 3-6
* *achievement*, grade-level average achievement test score
* *gap\_gender*, grade-level male-female gender gap on achievement test
* *percent\_ell*, district-level percentage of ELL students
* *percent\_sped*, district-level percentage of students in special education program
* *percent\_frl*, district-level percentage of students eligible for free or reduced school lunch
* *percent\_native*, district-level percentage of Native American students
* *percent\_asian*, district-level percentage of Asian students
* *percent\_hispanic*, district-level percentage of Hispanic students
* *percent\_black*, district-level percentage of African American students
* *percent\_white*, district-level percentage of white students