Crimes

The crimes data contain information on crime rates and demographic variables for all fifty US states and the District of Columbia. The crime rates are averaged over the years 2014-2018 and come from the FBI's Uniform Crime Reports (UCR). The UCR is a program where local law enforcement agencies all report crime statistics to the FBI and these are aggregated into final crime statistics. For our purposes, we are dividing crimes into two main categories of violent and property crime.

The demographic characteristics come from the American Community Survey (ACS) between the years 2014 and 2018. The ACS is an annual sample of the US population. To get a large enough sample in each state to calculate correct statistics (with little sampling error), I combine five years of data that are "centered"" on 2016.

Here is a full description of all variables in the dataset that we will use.

- **violent_rate**: violent crimes per 100,000 population within each state. This includes the crimes of murder, rape, robbery, and aggravated assault. By dividing the number of crimes by the population size, we avoid the problem of larger population states having more crimes because of a larger population. This is often called the crime "rate."
- **property_rate**: property crimes per 100,000 population. This includes the crimes of burglary, larceny, and motor vehicle theft.
- **median_age**: Median age of a state's population.
- **percent_male**: Percent of a state population that is male.
- **percent_lhs**: The percent of the state population over the age of 25 without a high school diploma.
- **median_income**: Median household income in a state. This is measured in thousands of 2018 US dollars (i.e. 35 means $35,000). We are taking the income of each household (meaning all members of that household combined) rather than individual level income. For most purposes, this is thought to be a better measure because consumption and savings are typically organized at the household level.
- **unemploy_rate**: Unemployment rate in the state. The unemployment "rate" is really just a percentage. Its the percentage of individuals who are not working but want to work among all those in the labor force (those who are working or looking for work).
- **poverty_rate**: Poverty rate in the state. The poverty "rate" is also really just a percentage. It is the percent of individuals living below the poverty line. The poverty line is a number developed by the federal government. It was originally developed in the 1960s and is adjusted for inflation every year. Many people critique the poverty line as being too low because it has not kept pace with increases in the consumer price index.
- **gini**: A measure of income inequality in the state. The gini coefficient is a widely used measure of how unequally income is distributed. If gini is

zero, then everyone has exactly the same income. If gini is 100, then one person makes all the money and everyone else zero. The higher the gini coefficient, the more income inequality exists.

Movies

The movie data contain information about 4,343 movies produced between 2000 and 2021. The data come from the Internet Movie Database and have been supplemented with extra information from the Open Movie Database. I have limited the total number of movies in the following ways:

- I have restricted the dataset to English language movies produced in the US (they may be filmed elsewhere).
- I have restricted movie runtime to movies that are at least 80 minutes long and no longer than 3.5 hours. The 80 minute benchmark is the lower limit for movies that the the Screen Actor's Guild considers "feature" films.
- I have restricted the dataset to movies that received at least 500 votes on the Internet Movie Database.
- I have restricted the dataset to movies that received a maturity rating between G and R.
- Movies must have valid responses on all variables in the Open Movie Database and must have made at least $100,000 domestically at the box office.
- I have excluded documentaries.

Here are the variables we have for each movie:

- **year**: The calendar year of the film's release.
- **runtime**: The length of the movie in minutes.
- **maturity_rating**: The movie's MPA maturity rating (G, PG, PG-13, or R).
- **genre**: The genre of the film. This is a tricky variable to create. In actuality, movies could be listed in up to three multiple genres in the IMDB. For example, "No Country for Old Men" is listed in the genres of crime, drama, and thriller while "Lord of the Rings: Return of the King" is listed as action, adventure, and fantasy. This is probably the best way to treat genres, but for our purposes it adds a lot of complexity. Therefore, I have recoded movies into a single "best" genre based on a decision rule where certain genres trump all others on an ordered basis. For example, comedy trumps romance, so romantic comedies will always show up in this dataset as comedies. The ordering of this system is Animation > Family > Western > Biography > Musical > Horror > Sci-Fi/Fantasy > Comedy > Sport > Romance > Action > Thriller > Mystery > Drama > All Others. For the most part, this system works well, but you may notice some odd discrepancies for a few movies.

- **box_office**: Gross domestic (US only) box office returns for the movie in millions of US dollars. These are not adjusted for inflation.
- **Oscars**: The number of Oscar awards that the movie received. This includes Oscars that go to individual actors (leading and supporting), as well as more general awards (best screenplay, editing, cinematography, etc.), and best picture overall.
- **rating_imdb**: This is average score (between 1 and 10) for a movie provided by IMDB users.
- **metascore**: The movie's metascore rating from metacritic. The metascore is a curated weighted average of reviewer scores from a variety of sources.
- **awards**: The number of Oscar awards that this movied recieved.

Politics

This data comes from the 2016 American National Election Study (ANES). The ANES is a survey of the American electorate that is conducted every two years. The study collects information on a variety of political attitudes and voting behaviors. For our purposes, we are going to primarily look at respondent's vote for president and attitudes on three issues: (1) birthright citizenship, (2) gay marriage, and (3) global warming. The variables we will look at are:

- **brcitizen**: Respondents were asked whether they would support a proposal to change the US Constitution to remove birthright citizenship (citizenship automatically granted to individuals born in the US regardless of their parent's citizenship status). Respondents could either favor, oppose, or neither favor or oppose.
- **gaymarriage**: Respondents were asked for their position on gay marriage and were given the choices of "no legal recognition", "civil union (but no marriage)", "support gay marriage."
- **globalwarm**: A question on whether the respondent believes that anthropogenic global warming is happening. I constructed this variable from two separate questions. The first question asks whether respondents think that global warming has been happening with the options being that it "probably has" or "probably has not." The second question asks whether respondents thought that global warming was caused by human activity (either entirely or partially). I combine these into a single dichotomous variable where individuals either think the earth is warming from human activity or that it is not warming from human activity, where the latter category includes people who think it isn't warming at all and people who think it is warming but not because of human activity.
- **party**: The political party with which the respondent identifies. This does not necessarily mean that a respondent is officially registered with a given party.
- **relig**: The respondent's religion. This category is based on the combination of people's statement about the kind of services they typically attend

along with several non-exclusive yes/no questions about their religion (e.g. evangelical, Pentecostal, agnostic, atheist).

- **age**: The age of the respondent.
- **gender**: The respondent's self-reported gender, recorded as "Male","Female", or "Other."
- **race**: the racial identification of the respondent. Respondents could write in multiple races, but to keep it simple, we will combine the small number of individuals who reported multiple races with those who listed "Other" as their race.
- **educ**: The education of the respondent. This is recorded as an ordinal variable. The "Some college" response indicates individuals who have attended college (including 2-year programs) but have not earned a BA.
- **income**: The family income of the respondent in 1000s of dollars. Respondents did not give actual dollar amounts here but rather indicated which bracket of income (e.g. $20,000-30,000) they fell within. For the purposes of our class, I randomly select an actual value within this bracket for each respondent.
- **workstatus**: The work status of the respondent. Respondents could either be working, unemployed, or out of the labor force. The last category refers to people who are not employed and not currently looking for work, whereas unemployed indicates a person who is not employed an is currently looking for work.
- **military**: Whether the respondent has ever served or is currently serving in the US military.

Popularity

This data comes from the National Longitudinal Study of Adolescent to Adult Health (Add Health), conducted by the Carolina Population Center at UNC-Chapel Hill and supported by a grant from the National Institute of Child Health and Human Development. The first wave of the study which we are using surveyed adolescents between 7th and 12th grade in school in the 1994-95 school year. One of the particularly valuable features of the Add Health survey is that many respondents were in the "saturation sample" which sampled all students at 16 schools. In this saturation sample, students were asked about who were their friends and sexual partners, which allows researchers to construct network maps of adolescent social systems.

We will use this saturation sample to look at a various basic measure of that network that estimates students' popularity. This measure, which is called "in degree" in the network analysis literature, measures the number of times a student was nominated as a friend by other students in the school. We will treat it as a simple proxy measure of a student's popularity. We can then look at what other student characteristics were positively or negatively associated with a student's popularity.

Here is a full description of all variables in the dataset that we will use.

- **grade**: Student's grade in school.
- **race**: A six-category nominal variable indicating the race that the student best thought described them when asked to choose a single race: white, black, Latino, Asian, American Indian, other.
- **gender**: Student's gender. Students were only reported as male or female.
- **nominations**: The number of friend nominations received by other students at the same school. This is the measure of popularity that we will use.
- **alcoholuse**: Students who reported drinking at least once or twice a month in the last twelve months were treated as "Drinkers" and all other students as "Non-drinkers.""
- **smoker**: Students who reported smoking more than 5 cigarettes in the past 30 days were treated as "Smokers" and all others as "Non-smokers."
- **pseudo_gpa**: Students were asked for the most recent letter grade in four course types: math, language arts, science, and math. This variable was constructed by calculating GPA from those four responses.
- **honor_society**: Whether the student was in honor society or not. Recorded as "Yes" or "No."
- **bandchoir**: Whether the student was in band or choir. Recorded as "Yes" or "No."
- **nsports**: The number of different school sports a student reported participating in. Students who reported more than six sports were top-coded at the value of six.
- **parent_income**: Parent's household income measured in $1000's of dollars.

Sex

The sex data come from the General Social Survey (GSS) for the years between 2014 and 2021. The GSS is a survey of attitudes that is conducted every two years by the National Opinion Research Council (NORC). In addition to many other questions, respondents were asked a question about the frequency of sexual activity. We will examine that variable as well as several other social and demographic characteristics and its relationship to demographic characteristics such as age, education, and marital status. Here are the variables we will look at:

- **sexf**: A quantitative variable indicating the frequency of sexual activity as the number of sexual encounters per year. The sexual frequency response was originally coded as an ordinal scale variable in which respondents were given a set of options from less to more sexual activity in the previous year. For our purposes, In order to have more quantitative data to work with, I have recoded this ordinal variable into a quantitative variable by

randomly assigning everyone a value around the mean of their ordinal response. This creates more noise in the dataset but should produce results that are generally consistent with the original ordinal scale.
- **gender**: The gender of the respondent.
- **age**: The age of the respondent. The GSS only surveys adults aged 18 years and older.
- **marital**: Marital status of the respondent: Never married, married, divorced, and widowed. A small number of "married, but separated" individuals are treated as "divorced" here.
- **sexorient**: Sexual orientation of the respondent: heterosexual, gay or lesbian, or bisexual.
- **relig**: Religious affiliation of the respondent. Protestants have been divided into "Mainline" and "Evangelical" based on a coding of specific denominations used by the GSS.
- **educ**: Years of education for the respondent.

Titanic

The titanic data contain information on all 1,309 passengers aboard the Titanic. The data do not include information about the crew. The data primarily come from the online database, Encyclopedia Titanica. Here are the variables we will look at:

- **survival**: Did the passenger survive?
- **sex**: The reported sex of the passenger.
- **age**: The age of the passenger. This variable is reported in whole numbers for those over one year old and as a decimal (based on months of age) for infants under a year of age.
- **agegroup**: A categorical variable indicating whether the person was an adult or a child. I have constructed this variable from the age variable. The cutoff for adults is sixteen years of age.
- **pclass**: There were three passenger classes: First, second, and third (also known as steerage). To give some pop culture references, Rose was first class, and Jack was third class. Most of the passengers were in third class.
- **fare**: The fare paid for the ticket, measured in British pounds.
- **family**: The number of family members traveling with the passenger. These family members can either be parents, spouses, siblings, or children.

Earnings

This data has information on the hourly wages of US workers in 2018. The data here are extracted from Current Population Survey data via IPUMS. I used the earning data from the outgoing rotation groups (ORG) for each month of the CPS. Each household in the CPS is is part of a rolling panel in which they are

in for four months, out for eight months, and back in for four months. In the fourth and eight month of inclusion they are given additional questions as part of the outgoing rotation group. The hourly wage of salaried workers is assessed by a question on hours worked in a typical week and earnings in the prior week.

I limited the data only to those individuals between the ages of 18 and 65 in order to capture the age range of the typical worker. The dataset contains the following variables:

- **wages**: The hourly wage for the respondent. For workers who report being paid hourly, this value is based on a direct question that asked for respondents' hourly wages. For individuals in salaried positions, this value was derived by dividing the earnings from the previous week by the hours worked in the previous week. Anyone who reported a wage of less than one dollar is removed. Any wage higher than $99.99 is top-coded as $99.99.
- **age**: age of the respondent in years.
- **gender**: Male or Female.
- **race**: The respondent's racial identification recoded from two separate questions on race and hispanicity into the following categories: White, Black, Latino, Asian, Indigenous, and Other/Multiple races. The indigenous category includes American Indians, Pacific Islanders, and Alaska Natives.
- **marstat**: The respondent's current marital status: never married, married, divorced or separated, and widowed.
- **education**: The respondent's highest educational attainment: no high school diploma, high school diploma, associate's degree, bachelor's degree, graduate degree. The last category includes master's degrees, professional degrees, and doctoral degrees.
- **occup**: The broad occupational category of the respondent. In the actual CPS data, there are hundreds of different occupations listed. For our purposes, I have simplified this into a broader (and smaller) set of occupational categories that we will use for the analysis. Here are the categories of the occupational variable, along with some examples of specific occupations:
  - *Managers*: Human resources Managers, Operations Managers
  - *Business/Finance Specialist*: Claims Adjusters, Compliance Officers, Accountants, Tax Preparers
  - *STEM*: Computer Programmers, Civil Engineers, Biological Scientists
  - *Doctors*: Dentists, Surgeons, Optometrists
  - *Legal*: Lawyers, Judges, Paralegals
  - *Education*: Preschool and Kindergarten Teachers, Librarians
  - *Arts, Design, and Media*: Artists, Dancers and Choreographers, Writers and Authors
  - *Other Healthcare*: Registered Nurses, Physical Therapists, Dental Hygienists
  - *Social Services*: Clergy, Social Workers
  - *Service*: Waiters and Waitresses, Barbers, Bartenders

  - *Sales*: Cashier, Telemarketer
  - *Administrative Support*: Bank Tellers, Data Entry Keyers, Receptionist
  - *Manual*: Carpenters, Logging Workers, Mining Machine Operators, Small Engine Mechanic

- **nchild**: Number of own children living in the household with the respondent.
- **foreign_born**: A variable indicating whether the respondent is foreign born or not. Recorded as "Yes" or "No".
- **earn_type**: This variable indicates whether the respondent reported being paid hourly wages or by salary.
- **earningwt**: A technical weighting variable for use with any CPS analysis of earnings.