Game Plan for Accessing Data Lecture Examples

* **General Social Surveys (GSS)**
  + Navigate to GSS Explorer
    - [GSS Data Explorer | NORC at the University of Chicago](https://gssdataexplorer.norc.org/MyGSS)
    - Encourage to make an account
  + Search variables
    - Be mindful of years/ballots question is asked
    - In example, find “natcrime” “courts” “fear” “sex” and “age”
    - Click “Add to MyGSS”
  + Download data
    - Navigate to your MyGSS account
    - Check boxes for above variables, include “year” and “ballot”
    - Under “Actions” click “Add to Extract”
    - Navigate to “Extracts” tab, click “Create Extract” then follow prompt to name, build, and download the extract
      * For output options, I suggest Excel (data +metadata). The R Script is not the actual data, just a script to transfer it from dictionary and data file.
    - Click “Create Extract” when finished
      * Page may hang, refresh and the extract should be processed
      * Actions -> Download
        + The file has an overly complicated structure and requires multiple rounds of unzipping subfolders.
  + Bring Data into R
    - Import from Excel, explore the data a little to show what it looks like (e.g., missing values, number of observations, structure of data, etc…)
* **Global Terrorism Database (GTD)**
  + Navigate to Canvas
    - <https://uncc.instructure.com/courses/187459/files/folder/Data%20Set%20Resources/Global%20Terrorism%20Database>
  + Bring data into R
    - Import from Excel
  + Explore data
    - Show what it looks like, missing values, number of observations, structure of data
      * Show what is contained in the country\_txt and iyear variables, plot both
  + Introduce codebook
    - Many variables in GTD, codebook tells you what they mean
    - When in doubt, check the codebook
      * Codebooks are user manuals for datasets – would you use something you have no clue about without reading the manual?
        + The answer is probably yes, but in the realm of power tools that’s a good way to lose a finger. Read the codebook.
* **National Longitudinal Study of Adolescent to Adult Health (Add Health)**
  + Navigate to Canvas
    - Download zipped folder here: <https://uncc.instructure.com/courses/187459/files/folder/Data%20Set%20Resources/Add%20Health>
    - Unzip using the proper program on your computer
  + Bring data into R
    - There are many DS00## subfolders within the main file
      * Each contains a different wave/type of data
      * Look at data documentation FIRST to see which file you want to import.
        + You will probably want the Wave “X” files, where X is a roman numeral that indicates which wave the data are from
    - Working with Wave I data today
      * Data are already in .RData format. You can simply double click on the file to have it open in RStudio
  + Explore data
    - Double-click on data frame to see in Console pane
      * Columns have descriptive labels underneath
        + Very helpful for remembering what data the variable contains
      * You will want to store the data frame under a different name. The provided one is long (you do not want to type that any more than once).
    - Take closer look at drug variables
      * Age first tried marijuana (H1TO30)
        + Variable values are strange here (multiple parentheses in them) and are stored as factor values.
        + Helpfully, NAs are already recorded
        + Plot variable (plot is messy – consider numeric)
      * Carry weapon to school (H1FV9)
        + Another factor variable with strange , long values. Makes some sense here, as the categories are not strictly numeric.
  + Introduce codebook
    - In addition to data guide, there is a codebook with variable values, frequencies, and text of the actual questions asked
    - My suggestion – Ctrl+F search for variables you want, record their names, then make a new data frame with just those variables in it.
      * May also want to make a separate word document with the variable description from the PDF file.
* **National Longitudinal Survey of Youth 1997 (NLSY97)**
  + Navigate to NLS Investigator Website
    - Here: [NLS Investigator Search (nlsinfo.org)](https://www.nlsinfo.org/investigator/pages/search)
      * Select NLSY97 as the “study you want to work with” then NLSY97 1997-2019 (rounds 1-19) as the substudy
    - I strongly suggest making an account on this site if you want to use these data
  + Searching for variables
    - Click on the “Variable Search” tab
      * You can browse the index or search for particular variables within the other two tabs
    - Go look within “Crime & Substance Use”
      * Various self-report criminal behavior variables within “Criminal activity” – explore this section
      * Within Substance use -> Illegal Drugs there are multiple other indicators for specific types of drug use.
    - ALWAYS grab the interview year variable
      * Go to “Survey & Methodology” -> “Interview date” -> “Month & Year
      * Expand the row, then click blank box for year
    - If you hover over the variable name/question name it will provide a preview of the question and frequency of different answers
    - Click the blank box to add the variable to your “cart”
  + Downloading data
    - Go to “Review Selected Variables” to make sure you have everything you wanted
      * If everything looks good, go to “Save/Download”
      * Navigate to “Advanced Download”
        + Uncheck all but “R Source Code”, “Codebook of selected variables”, “Short description file”, and “Comma-delimited datafile”

The codebook is a .cdb file for some unearthly reason. You can open it in a notepad program.

The short description file can also be opened in a notepad program.

* + - * Provide name for the datafile in the “Filename” box, then click “Download”
        + Click to “download” again on the next page

Will also download as a zipped folder

* + Bringing data into R
    - Need to use the “From text (base)” option
      * Make sure that the separator is set to “Comma”, “Heading” is set to “Yes”, and the name of the file is what you want it to be
  + Explore data
    - NLSY97 data are messy.
      * Every variable has been imported as integer class but that’s not accurate.
      * The data has also been imported with several different types of missing data values (values <0). These need to be recoded properly.
        + Generally, -4 stands for “Valid Skip” and means that the youth responded “No” to some lead-in question. You can usually code these as 0s for frequency variables and 0s for Yes/No variables (but always check codebook!).
        + The remaining missing values correspond to Invalid Skips (youth skipped without answering), non-interviews, and “Don’t Know” responses.
    - Be mindful of time question was asked
      * Be sure to clearly label variables with their years if you plan to use values from the same question for multiple years.
        + For example, marijuana use in the past 30 days may be called marj\_freq97 for the 1997 interview, marj\_freq98 for the 1998 interview, and so on.

Come up with a system, and stick to it.

* + - Explore marijuana use in past 30 days variable and year of birth.
* **Pathways to Desistance (PtD)**
  + Navigate to Canvas
    - Data file is stored here: <https://uncc.instructure.com/courses/187459/files/folder/Data%20Set%20Resources/Pathways%20to%20Desistance>
    - This is a zipped version of the baseline survey results and all follow-ups. There are 11 total waves of data within this folder.
  + Bring data into R
    - Each of the DS00## files corresponds to an individual wave of data collection, beginning with the baseline interview (DS0001) and ending with the Wave 20 interview (DS0011).
    - These files were also downloaded in a format compatible with R, so you should be able to simply double click on the file for it to open in RStudio.
  + Explore the data
    - Weinberger Adjustment Inventory – Impulse Control (S0IMPULS)
      * They’ve been so kind as to code the missing for us already – this may be a good thing but depends on the question.
    - Personal Rewards of Crime (S0PERRWD)
      * Measures the subjective personal rewards to committing crime (“fun or kick”) on a scale of 1 (no fun) to 10 (maximum fun)
  + Introduce the codebook
    - There’s one codebook for every file.
    - Some questions may change slightly in some waves (uncommon) or be introduced at later waves (also uncommon)
      * Good example is the wage information – they initially only collected information about legal/illegal income, but then started asking questions about “under the table” work at a later wave
    - Minor gripe – getting the full text for a question is tricky
      * Many measures are from large psychometric scales that you can find published elsewhere (an extra step) and a small amount of measures are from proprietary scales (which you’d have to pay for to get exact question text)