MITLibraries

Understanding, Finding, and Using Data

17.871

Spring 2012

MITLibraries

Goals for Today

- Understanding research datasets
- Identifying potential data sources
- Hands-on exercises to download data for your upcoming problem set
- Searching for research data from ICPSR

MITLibraries

Today's Research Topics

1. The percentage of people with driver's licenses, by age and state
2. Patterns of conflicts among countries

- How can we find the right data? Look for:
  - Data on certain topics
  - During certain time periods
  - In particular geographic areas
  - Data can be challenging to find—start early

MITLibraries

Social Science Data Services

- Support for finding and managing data
- Libraries' research guide: http://libraries.mit.edu/guides/subjects/data
  - Data Access
  - Training
  - Hardware & Software
- GIS Services: http://libraries.mit.edu/gis

Data File Structure

- Fixed-field vs. delimited
- Rectangular/LRECL vs. card image vs. hierarchical
- Unit of analysis (e.g., person, household, administrative unit, event)

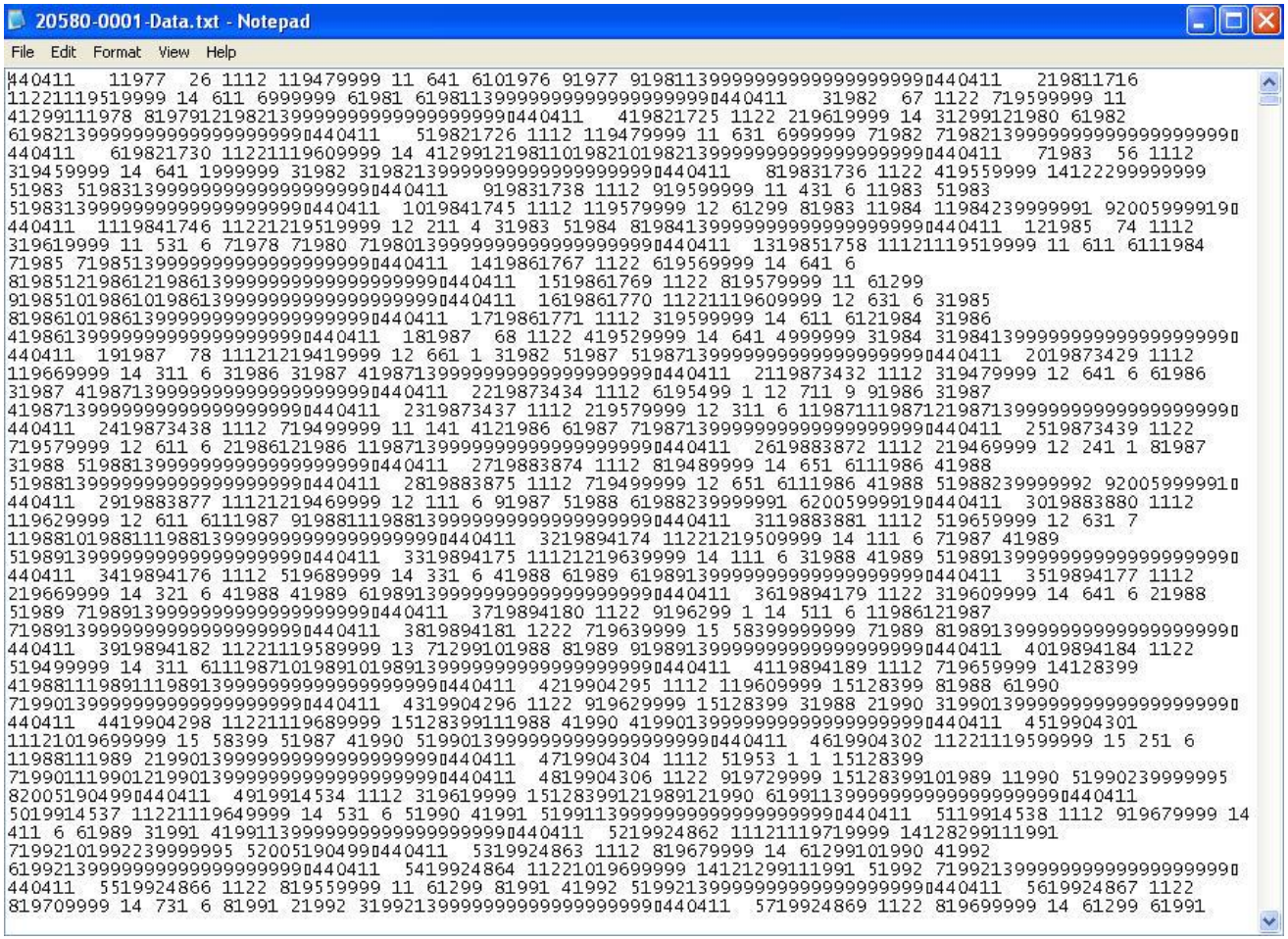
MITLibraries

Types of Variables

- Alpha/character vs. numeric
- Continuous vs. discrete/categorical
- Micro-level vs. summary-level/aggregate
- Weight variables
- Tutorials on Understanding Data Files: http://libraries.mit.edu/guides/subjects/data/training/tutorials/understanding.html

MITLibraries

Sample Data File


20580-0001-Data.txt - Notepad

File Edit Format View Help

```
440411   11977  26 1112 119479999 11 641 6101976 91977 919811399999999999999999999440411   219811716
11221119519999 14 611 6999999 61981 619811399999999999999999999440411   31982  67 1122 719599999 11
41299111978 819791219821399999999999999999999440411   419821725 1122 219619999 14 31299121980 61982
619821399999999999999999999440411   519821726 1112 119479999 11 631 6999999 71982 719821399999999999999999999
440411   619821730 11221119609999 14 41299121981101982101982139999999999999999999440411   71983  56 1112
319459999 14 641 1999999 31982 319821399999999999999999999440411   819831736 1122 419559999 14122299999999
51983 519831399999999999999999999440411   919831738 1112 919599999 11 431 6 11983 51983
519831399999999999999999999440411  1019841745 1112 119579999 12 61299 81983 11984 11984239999991 920059999919
440411  1119841746 11221219519999 12 211 4 31983 51984 819841399999999999999999999440411  121985  74 1112
319619999 11 531 6 71978 71980 719801399999999999999999999440411  1319851758 11121119519999 11 611 6111984
71985 719851399999999999999999999440411  1419861767 1122 619569999 14 641 6
819851219861219861399999999999999999999440411  1519861769 1122 819579999 11 61299
919851019861019861399999999999999999999440411  1619861770 11221119609999 12 631 6 31985
819861019861399999999999999999999440411  1719861771 1112 319599999 14 611 6121984 31986
419861399999999999999999999440411  181987  68 1122 419529999 14 641 4999999 31984 319841399999999999999999999
440411  191987  78 11121219419999 12 661 1 31982 51987 519871399999999999999999999440411  2019873429 1112
119669999 14 311 6 31986 31987 419871399999999999999999999440411  2119873432 1112 319479999 12 641 6 61986
31987 419871399999999999999999999440411  2219873434 1112 6195499 1 12 711 9 91986 31987
419871399999999999999999999440411  2319873437 1112 219579999 12 311 6 119871119871219871399999999999999999999
440411  2419873438 1112 719499999 11 141 4121986 61987 719871399999999999999999999440411  2519873439 1122
719579999 12 611 6 21986121986 119871399999999999999999999440411  2619883872 1112 219469999 12 241 1 81987
31988 519881399999999999999999999440411  2719883874 1112 819489999 14 651 6111986 41988
519881399999999999999999999440411  2819883875 1112 719499999 12 651 6111986 41988 51988239999992 920059999919
440411  2919883877 11121219469999 12 111 6 91987 51988 619882399999991 620059999191440411  3019883880 1112
119629999 12 611 6111987 919881119881399999999999999999999440411  3119883881 1112 519659999 12 631 7
119881019881119881399999999999999999999440411  3219894174 11221219509999 14 111 6 71987 41989
519891399999999999999999999440411  3319894175 11121219639999 14 111 6 31988 41989 519891399999999999999999999
440411  3419894176 1112 519689999 14 331 6 41988 61989 619891399999999999999999999440411  3519894177 1112
219669999 14 321 6 41988 41989 619891399999999999999999999440411  3619894179 1122 319609999 14 641 6 21988
51989 719891399999999999999999999440411  3719894180 1122 9196299 1 14 511 6 11986121987
719891399999999999999999999440411  3819894181 1222 719639999 15 58399999999 71989 819891399999999999999999999
440411  3919894182 11221119589999 13 71299101988 81989 919891399999999999999999999440411  4019894184 1122
519499999 14 311 611198710198910198913999999999999999999440411  4119894189 1112 719659999 14128399
419881119891119891399999999999999999999440411  4219904295 1112 119609999 15128399 81988 61990
719901399999999999999999999440411  4319904296 1122 919629999 15128399 31988 21990 319901399999999999999999999
440411  4419904298 11221119689999 15128399111988 41990 419901399999999999999999999440411  4519904301
11121019699999 15 58399 51987 41990 519901399999999999999999999440411  4619904302 11221119599999 15 251 6
11988111989 219901399999999999999999999440411  4719904304 1112 51953 1 1 15128399
719901119901219901399999999999999999999440411  4819904306 1122 919729999 15128399101989 11990 51990239999995
8200519049904404411  4919914534 1112 319619999 151283991219891219990 619911399999999999999999999440411
5019914537 11221119649999 14 531 6 51990 41991 519911399999999999999999999440411  5119914538 1112 919679999 14
411 6 61989 31991 419911399999999999999999999440411  5219924862 11121119719999 14128299111991
719921019922399999995 5200519049904404411  5319924863 1112 819679999 14 61299101990 41992
619921399999999999999999999440411  5419924864 11221019699999 14121299111991 51992 719921399999999999999999999
440411  5519924866 1122 819559999 11 61299 81991 41992 519921399999999999999999999440411  5619924867 1122
819709999 14 731 6 81991 21992 319921399999999999999999999440411  5719924869 1122 819699999 14 61299 61991
```

MITLibraries

Sample Codebook: TOC

TABLE OF CONTENTS

MITLibraries

Sample Codebook: Variables

Y2.

Are you married now and living with your (husband/wife) -- or are you widowed, divorced, separated, or have you never married?

Note: in cases 8,49,77,157,377,535,590,814,1087,1117,1202,1212, 1381,1639,1731 the partner/spouse identified in V000909 (marital status) was not a resident of the HU or was residing elsewhere on temporary basis. In cases 1291,1516,46,312,133,23,222 there is no information about spouse/partner.

1. MARRIED
2. WIDOWED
3. DIVORCED
4. SEPARATED
5. NEVER MARRIED
6. PARTNERED, NOT MARRIED [VOL]

8. DK
9. RP
0. NA

	0	1	2	3	4
Count	5	935	168	238	55

	5	6	8	9
Count	348	49	1	8

VAR 000910 Y3. Highest grade completed
MD1: EQ 99, MD2: GE 99
COLUMNS: 1813 - 1814
Numeric

Y3.

What is the highest grade of school or year of college you have completed?

00-12 YEARS
13-16 YEARS --> SKIP TO Y3b
17. 17+ YEARS --> SKIP TO Y3b

98. DK
99. RF; NA

	0	2	4	5	6	7
Count	3	3	5	2	14	12

MITLibraries

Identify Potential Data Sources

- Data is usually expensive and time-consuming to collect, store, and publish
- Who has the time, funds, and authority to collect the data?
- Why might someone want the data?
- Who is responsible for collecting or managing the data?
- Who might be external stakeholders?

MITLibraries

Searching for Data

- Data Access by Subject
- Data Centers
  - ICPSR
  - Harvard-MIT Data Center (HMDC)
- Tips on locating data
  - Search the political science literature

MITLibraries

Literature Searches

- Research guides in all fields: http://libraries.mit.edu/research-guides
- Worldwide Political Science Abstracts
- Google Scholar
- Citation Software at MIT: E.g., Refworks

Statistical Abstract of the U.S.

- Key source for identifying sources of data on various topics
- Contains tables of data and footnotes that link to sources of the data and more detail
- Search or browse
- Looking for: the number of licensed drivers by age group and state
- http://libraries.mit.edu/get/stat-abstract

MITLibraries

Exercise: Part 1

- Accessing data sources noted in the Statistical Abstract
- From table 1106 navigate to its original source
- In the original source of data, find and download the most recent year's data, broken down by age (as well as state)
- Write down for yourself the source of this data file you have just downloaded—be precise!

MITLibraries

Exercise: Part 2

- Find a table giving total population by state, from the 2010 Decennial Census, from American Fact Finder: http://factfinder2.census.gov/
- Use the Guided Search to select:
  - people > basic count > population total; age & sex > age
  - geographic type: state > all states
- Note the options of:
  - estimates from the ACS (American Community Survey)
  - data from the 2010 Census (2010 Demographic Profile)
  - *Profile of General Population and Housing Characteristics: 2010
- Download the table (+documentation) to your Athena locker as a .csv file (with or without annotations is fine)
- Write down the source of the data

Accessing Research Data: ICPSR

- World's largest social science data archive
- Collects and disseminates data sets
- Training in quantitative analysis
- Summer internship and research paper competition
- Note: responsible use and citing data
- MyData registration system
- Data also accessible via the Harvard-MIT Data Center: http://libraries.mit.edu/get/hmdc

ICPSR Demonstration

- http://libraries.mit.edu/get/icpsr
- Sample topic: patterns of conflicts among countries
- Main search
- Browsing by topic
- Reviewing a study
- Bibliography of Data-Related Literature
- Downloading data

Exercise: Part 3

- In ICPSR: http://libraries.mit.edu/get/icpsr
- Search for the following study: #24386, Correlates of War Project: Militarized Interstate Dispute (MID) Data, 1816-2001
- Download to your Athena directory:
  - DS2: MIDB: Participants in Each of the Disputes in MIDA (Version 3.10) (should come with all relevant documentation)

MITLibraries

Format Data

- Import the data into your statistical software package
- Can use setup files to import the data or Stat-transfer to convert among formats
- Extract variables of interest
- Subset observations of interest

Conclusion

- Feel free to contact me for help:

Katherine McNeill

mcneillh@mit.edu

- http://libraries.mit.edu/ask-us/
- Thanks for coming!
- More questions?