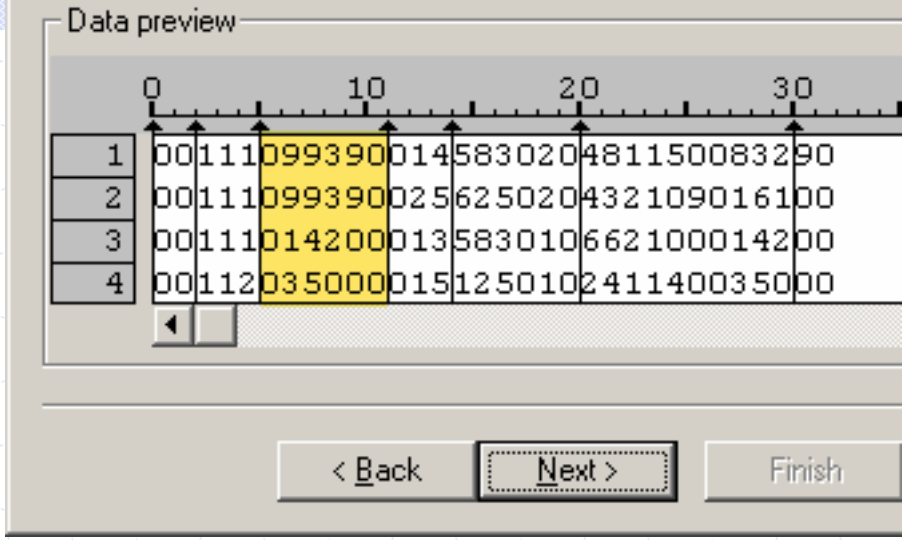
Demystifying Data Reference

Review: Data and Statistics

Raw Data



- Raw data: a file of numbers organized in rows and columns
- Each row has information about a particular unit: a person, a country…
- The columns represent *variables*, which are specific pieces of information about each unit
- In this census data, family income is in columns 6-11

Statistics are computed from data

- A data file may contain information about hundreds of thousands of units
- Statistical software is used to summarize this mess of numbers to produce usable information
- Here we've calculated that the average household income in this sample is around $73,000

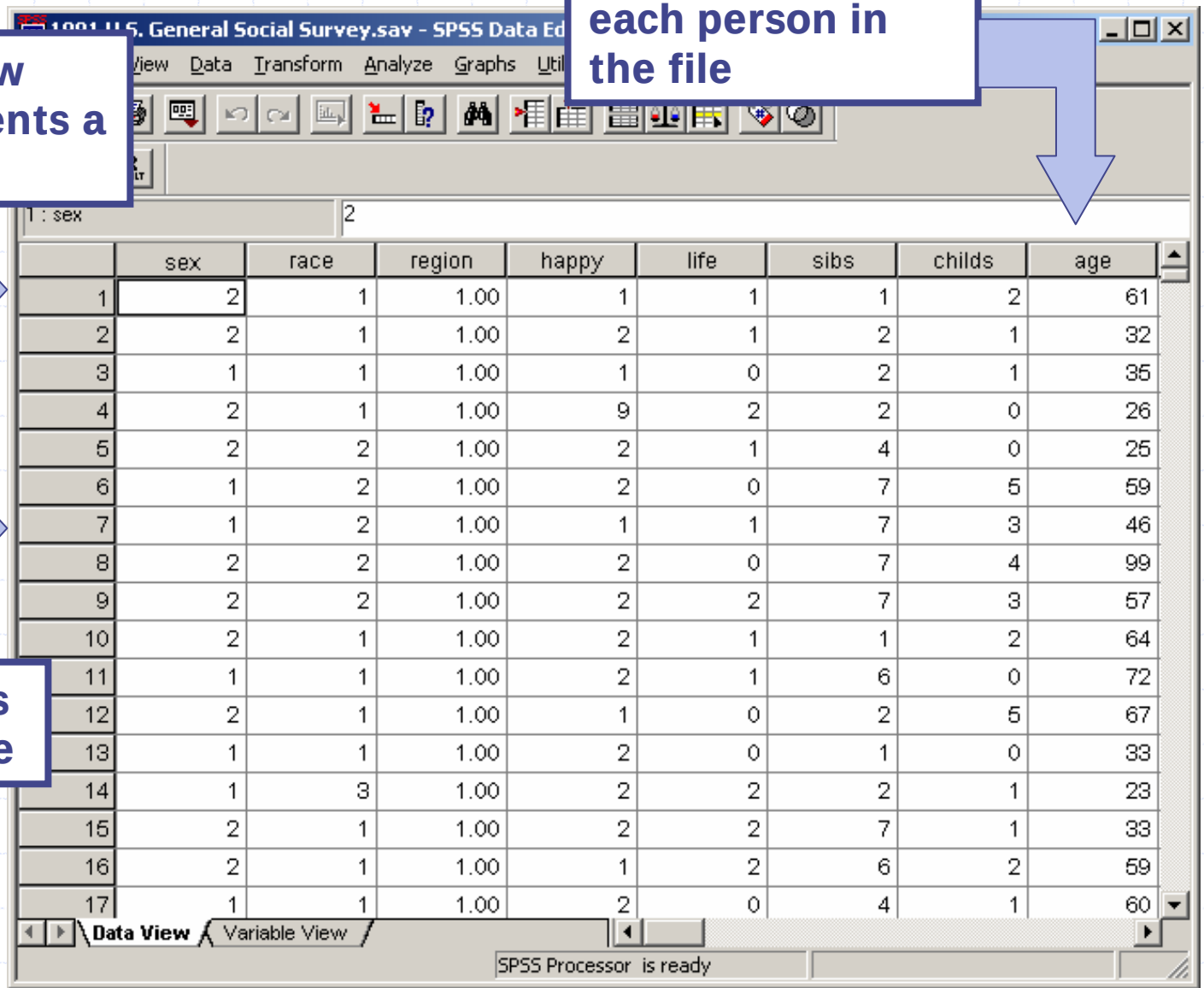
```
. sum hhincome

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    hhincome |    135276    73444.89     70013.4    -10000     999998

.
```

SPSS data view

This variable gives the age of each person in the file

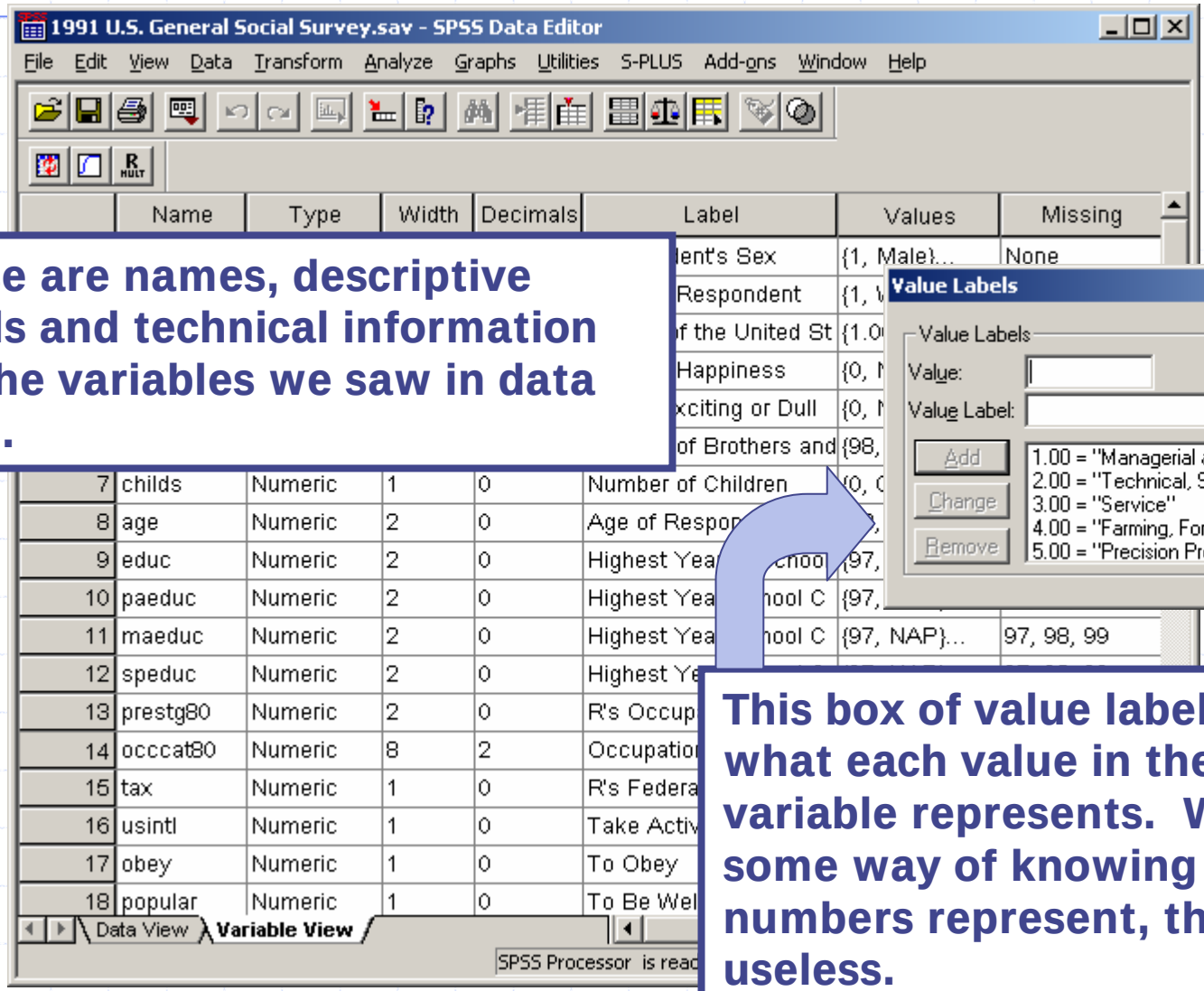
This row represents a person

So does this one



	sex	race	region	happy	life	sibs	childs	age
1	2	1	1.00	1	1	1	2	61
2	2	1	1.00	2	1	2	1	32
3	1	1	1.00	1	0	2	1	35
4	2	1	1.00	9	2	2	0	26
5	2	2	1.00	2	1	4	0	25
6	1	2	1.00	2	0	7	5	59
7	1	2	1.00	1	1	7	3	46
8	2	2	1.00	2	0	7	4	99
9	2	2	1.00	2	2	7	3	57
10	2	1	1.00	2	1	1	2	64
11	1	1	1.00	2	1	6	0	72
12	2	1	1.00	1	0	2	5	67
13	1	1	1.00	2	0	1	0	33
14	1	3	1.00	2	2	2	1	23
15	2	1	1.00	2	2	7	1	33
16	2	1	1.00	1	2	6	2	59
17	1	1	1.00	2	0	4	1	60

SPSS variable view



These are names, descriptive labels and technical information for the variables we saw in data view.

This box of value labels tells us what each value in the "occat80" variable represents. Without some way of knowing what the numbers represent, the data is useless.

Sample Codebook

```
Samples selected:
  2000 1% Census PUMS                                        small
File Type:                    rectangular
Compression:                   Yes
Case Selection:               None
  Variable    Columns Len 2000    Doc
  YEAR          1-  2    2     X
  GQ            3-  3    1     X
  FARM          4-  4    1     X
  OWNERSHP g    5-  5    1     X
  FTOTINC       6- 11    6     X
  PERNUM       12- 13    2     X
  PERWT        14- 17    4     X
  FAMSIZE      18- 19    2     X
  AGE          20- 22    3     X
  SEX          23- 23    1     X
  RACE      g  24- 24    1     X
  UHRSWORK     25- 26    2     X
  INCTOT       27- 32    6     X

Variable Availability Key:
All Years X - available in this sample
All Years . - not available in this sample

Case Selections:
```

Types of data

Navigating the terminology

Macro vs. Micro

Microdata is about individuals, macrodata is about populations

Macro data

- Macro data is country, state or region level data such as employment rate, GDP, infant mortality, etc.
  - Time Series: one country/unit over time
  - Cross-sectional: multiple countries
  - Longitudinal/Panel: both

Micro data

- Micro data is data on individual people or units, such as households, families, stocks or firms
- Reasons to use micro data:
  - Aggregates you need aren't available, or aren't available broken down in the way you want
  - Want to conduct analysis of relationships between different individual characteristics

Survey data – demographics and opinions

- Most micro data comes from surveys
- A good first question to ask when looking for survey data is "Who cares about what I am studying?"
- Unfortunately, the answer may be “No one.”

Demographic Survey Data

- Mostly collected by government agencies
- Facts about individual people, families or households – age, income, length of residency, drug use, age of third child, etc.
- Micro data collected from economic or demographic surveys and censuses
  - **Not** census counts – those are macro!
  - Look at individual-level census data, economic surveys such as the National Census, Current Population Survey, Survey of Income and Program Participation, etc.
  - Data collected by government agencies is often purely demographic

Examples of Demographic Data

- Census PUMS
  - http://www.ipums.org/
- Current Population Survey
  - http://beta.ipums.org/cps
- Demographic and Health Surveys
  - http://www.measuredhs.com/

Opinion Survey Data

- Data on people's *ideas* or *opinions*
- Look at social surveys and public opinion polls

Academic Social Surveys

- Large scale social science surveys ask questions about basic attitudes, opinions and values, broad trends in society
- Often also include relatively detailed demographic information

Major Academic Social Surveys

- General Social Survey
- World Values Survey
  - Both at ICPSR
- American National Election Studies
  - http://www.umich.edu/~nes/

Public Opinion Polls

- Generally include only a few demographic questions – often just age, sex, race, education and income.
- Opinion polls contain reactions to specific events, snapshots of opinion at particular moments in time on “hot issues”
  - Gallup, Media Surveys (CBS News), the Barometer series (Eurobarometer, Afrobarometer)
  - http://people-press.org/dataarchive/

Longitudinal Survey Data

- Data that follows the **same** people or units over time.
  - Use to study how people change over time
  - Look for things with “panel” or “longitudinal” in the title or description
    - National Educational Longitudinal Study, Panel Study of Income Dynamics (PSID), National Longitudinal Survey of Youth
  - For following general trends over time, **cross sectional** social surveys that are repeated regularly on different samples are equally useful, and easier to find.

Things to ask about

- Unit of analysis – person? School? State?
  - Macro or Micro
- Time period – how current? Single year or multiple years?
- Geographic coverage – multinational, national, regional
- Sample characteristics