Course > EDA: Examining Distributions > Exploratory Data Analysis (EDA) Overview > Data and Variables

Bookmark this page

Data and Variables

Before we jump into exploratory data analysis and really appreciate its importance in the process of statistical analysis, let's step back for a minute and ask:

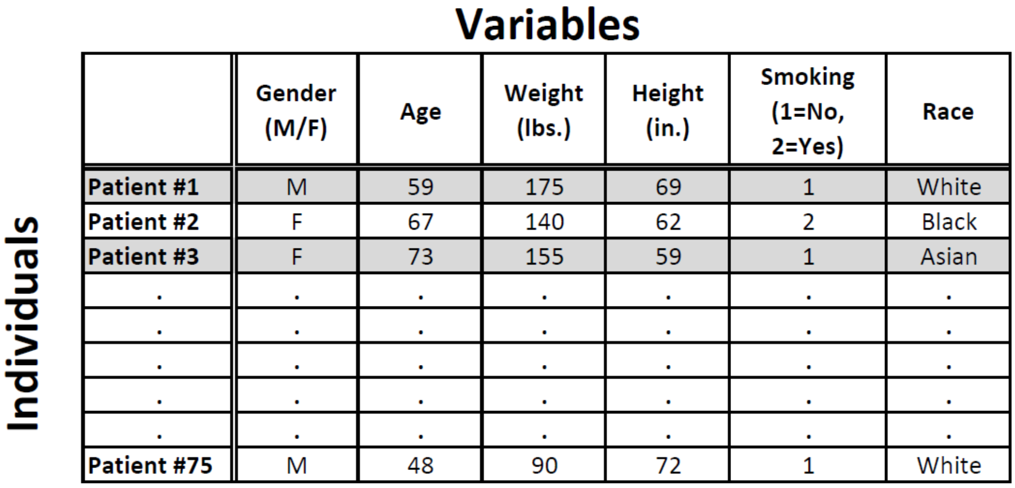
What do we really mean by *data*?

Data are pieces of information about individuals organized into variables. By an **individual**, we mean a particular person or object. By a **variable**, we mean a particular characteristic of the individual.

A **dataset** is a set of data identified with particular circumstances. Datasets are typically displayed in tables, in which rows represent individuals and columns represent variables.

Example: Medical Records

The following dataset shows medical records from a particular survey:


Variables
Individuals

	Gender (M/F)	Age	Weight (lbs.)	Height (in.)	Smoking (1=No, 2=Yes)	Race
Patient #1	M	59	175	69	1	White
Patient #2	F	67	140	62	2	Black
Patient #3	F	73	155	59	1	Asian
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
Patient #75	M	48	90	72	1	White

In this example, the individuals are patients, and the variables are Gender, Age, Weight, Height, Smoking, and Race. Each row, then, gives us all the information about a particular individual (in this case, patient), and each column gives us information about a particular characteristic of all the patients.

Variables can be classified into one of two types: categorical or quantitative.

- **Categorical variables** take category or label values and place an individual into one of several groups. Each observation can be placed in *only* one category, and the categories are mutually exclusive.

  In our example of medical records, Smoking is a categorical variable, with two groups, since each participant can be categorized only as either a nonsmoker or a smoker. Gender and Race are the two other categorical variables in our medical records example. (Notice that the values of the categorical variable Smoking have been coded as the numbers 1 or 2. It is common to code the values of a categorical variable as numbers, but you should remember that these are just codes. They have no arithmetic meaning (i.e., it does not make sense to add, subtract, multiply, divide, or compare the magnitude of such values).

- **Quantitative variables** take numerical values and represent some kind of measurement.

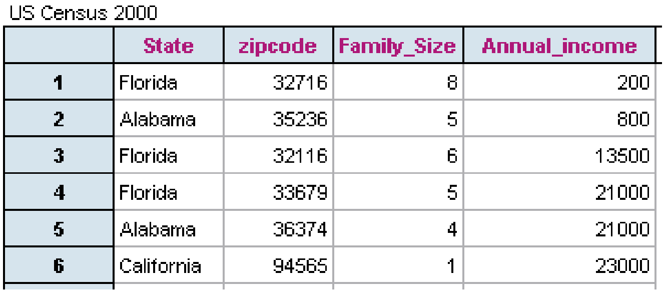
  In our medical example, Age is an example of a quantitative variable because it can take on multiple numerical values. It also makes sense to think about it in numerical form; that is, a person can be 18 years old or 80 years old. Weight and Height are also examples of quantitative variables.

NOTE...

Categorical variables are sometimes called qualitative variables, but in this course we use the term categorical.

Scenario: U.S. Census

We took a random sample from the 2000 U.S. Census. Here is part of the dataset:



US Census 2000

	State	zipcode	Family_Size	Annual_income
1	Florida	32716	8	200
2	Alabama	35236	5	800
3	Florida	32116	6	13500
4	Florida	33679	5	21000
5	Alabama	36374	4	21000
6	California	94565	1	23000

Learn By Doing

1/1 point (graded)

Who are the individuals described by this data?

○ States

● People living in the United States in the year 2000 ✓

○ People with families in the year 2000

Answer

Correct: The U.S. Census is completed by people living in the United States.

Submit

Learn By Doing

1/1 point (graded)

What type of variable is Zipcode?

● Categorical ✓

○ Quantitative

Answer

Correct: Zipcode is a categorical variable because it categorizes individuals by geographic location.

Submit

Learn By Doing

1/1 point (graded)

What type of variable is Family Size?

○ Categorical

● Quantitative ✓

Answer

Correct: Family size is a variable with numerical values that can be averaged.

Submit

Learn By Doing

1/1 point (graded)

What type of variable is Annual Income?

○ Categorical

● Quantitative ✓

Answer

Correct: Annual income is a variable with numerical values that can be averaged.

Submit

Clinical Depression and Drug Treatment

Background

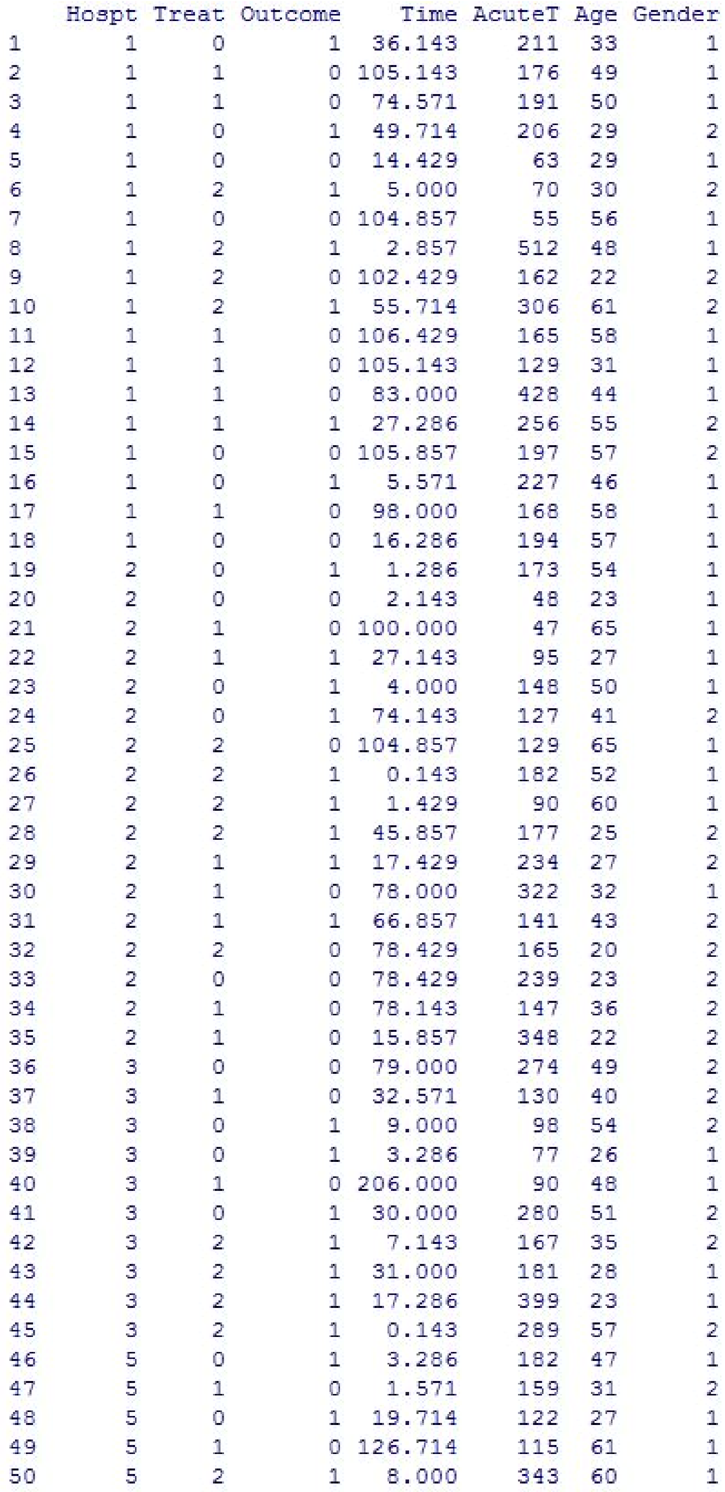
Clinical depression is the most common mental illness in the United States, affecting 19 million adults each year (Source: NIMH, 1999). Nearly 50% of individuals who experience a major episode will have a recurrence within 2 to 3 years. Researchers are interested in comparing therapeutic solutions that could delay or reduce the incidence of recurrence.

In a study conducted by the National Institutes of Health, 109 clinically depressed patients were separated into three groups, and each group was given one of two active drugs (imipramine or lithium) or no drug at all. For each patient, the dataset contains the treatment used, the outcome of the treatment, and several other interesting characteristics.

Here is a summary of the variables in our dataset:

- **Hospt:** The patient's hospital, represented by a code for each of the 5 hospitals (1, 2, 3, 5, or 6)
- **Treat:** The treatment received by the patient (Lithium, Imipramine, or Placebo)
- **Outcome:** Whether or not a recurrence occurred during the patient's treatment (Recurrence or No Recurrence)
- **Time:** Either the time in days till the first recurrence, or if a recurrence did not occur, the length in days of the patient's participation in the study.
- **AcuteT:** The time in days that the patient was depressed prior to the study.
- **Age:** The age of the patient in years, when the patient entered the study.
- **Gender:** The patient's gender (1 = Female, 2 = Male)

Here's a snapshot of the first 50 patients in the dataset with gender recoded to display Female or Male:



```
    Hospt Treat Outcome    Time AcuteT Age Gender
1       1     0       1  36.143    211  33      1
2       1     1       0 105.143    176  49      1
3       1     1       0  74.571    191  50      1
4       1     0       1  49.714    206  29      2
5       1     0       0  14.429     63  29      1
6       1     2       1   5.000     70  30      2
7       1     0       0 104.857     55  56      1
8       1     2       1   2.857    512  48      1
9       1     2       0 102.429    162  22      2
10      1     2       1  55.714    306  61      2
11      1     1       0 106.429    165  58      1
12      1     1       0 105.143    129  31      1
13      1     1       0  83.000    428  44      1
14      1     1       1  27.286    256  55      2
15      1     0       0 105.857    197  57      2
16      1     0       1   5.571    227  46      1
17      1     1       0  98.000    168  58      1
18      1     0       0  16.286    194  57      1
19      2     0       1   1.286    173  54      1
20      2     0       0   2.143     48  23      1
21      2     1       0 100.000     47  65      1
22      2     1       1  27.143     95  27      1
23      2     0       1   4.000    148  50      1
24      2     0       1  74.143    127  41      2
25      2     2       0 104.857    129  65      1
26      2     2       1   0.143    182  52      1
27      2     2       1   1.429     90  60      1
28      2     2       1  45.857    177  25      2
29      2     1       1  17.429    234  27      2
30      2     1       0  78.000    322  32      1
31      2     1       1  66.857    141  43      2
32      2     2       0  78.429    165  20      2
33      2     0       0  78.429    239  23      2
34      2     1       0  78.143    147  36      2
35      2     1       0  15.857    348  22      2
36      3     0       0  79.000    274  49      2
37      3     1       0  32.571    130  40      2
38      3     0       1   9.000     98  54      2
39      3     0       1   3.286     77  26      1
40      3     1       0 206.000     90  48      1
41      3     0       1  30.000    280  51      2
42      3     2       1   7.143    167  35      2
43      3     2       1  31.000    181  28      1
44      3     2       1  17.286    399  23      1
45      3     2       1   0.143    289  57      2
46      5     0       1   3.286    182  47      1
47      5     1       0   1.571    159  31      2
48      5     0       1  19.714    122  27      1
49      5     1       0 126.714    115  61      1
50      5     2       1   8.000    343  60      1
```

Did I Get This

1/1 point (graded)

Who are the individuals described by this data?

○ 19 million adults who experience depression each year

○ Hospitals

● 109 clinically depressed people ✓

Answer
Correct:
The dataset contains information on 109 clinically depressed people who were part of the NIH study.

Submit

Did I Get This

1/1 point (graded)
Which of the following variables is categorical? Check all that apply.

☑ Treat

□ AcuteT

☑ Outcome

✓

Answer
Correct:
Treat and Outcome are both categorical variables. Treat is a categorical variable because the treatment received by the patients is in the form of categories (Lithium, Imipramine, or Placebo). Outcome is a categorical variable since recurrence is in the form of two categories (Recurrence or No Recurrence).

Submit

Did I Get This

1/1 point (graded)
Which of the following variables is quantitative? Check all that apply.

□ Hospt

☑ Time

☑ Age

□ Gender

✓

Answer

Correct:

Time and Age are quantitative variables, since they can take on multiple numerical values, which have arithmetic meaning (i.e., it makes sense to add, subtract, multiply, divide, or compare the magnitude of such values).

Submit