Data Basics

David Gerard

2017-09-18

Learning Objectives

- Cases/observational units
- Variables (categorical/quantitative)
- Data frames
- Section 1.2 of DBC

Cases and Variables

- At its most basic level, data consist of two things: cases and variables.

case

Cases (or observational units) are the objects described by a set of data. Cases may be customers, companies, subjects in a study, units in an experiment, or other objects.

variable

For each case, the data give values for one or more variables. A variable describes some characteristic of a case, such as a person's height, gender, or salary. Variables can have different values for different cases.

Trump's Twitter

A data frame giving characteristics of President Trump's tweets from 2015-12-14 to 2016-08-08.

Subset of variables:

- `source` Whether the tweet came from an Android or an iPhone.
- `text` The text of the tweet.
- `hour` The hour of the day the tweet was created, from 0 to 23.
- `length` The length of the tweet.
- `anger` Whether the tweet has a word in it that evokes anger.
- `negative` Whether the tweet has a word in it that evokes a negative sentiment.

Cases? Variables?

```
library(tidyverse) ## for %>%, select, and glimpse
read_csv("../../data/trump.csv") %>%
  select(source, text, hour,
         length, anger, negative) ->
  trump
glimpse(trump)
```

```
## Observations: 1,390
## Variables: 6
## $ source   <chr> "Android", "iPhone", "iPhone...
## $ text     <chr> "My economic policy speech w...
## $ hour     <int> 10, 8, 19, 18, 16, 8, 21, 21...
## $ length   <int> 67, 90, 40, 134, 135, 138, 5...
## $ anger    <lgl> FALSE, FALSE, FALSE, TRUE, T...
## $ negative <lgl> FALSE, FALSE, FALSE, TRUE, T...
```

fumbles dataset

This data frame gives the number of fumbles by each NCAA FBS team for the first three weeks in November, 2010.

A data frame with 120 observations on the following 7 variables.

- `team` NCAA football team
- `rank` rank based on fumbles per game through games on November 26, 2010
- `W` number of wins through games on November 26, 2010
- `L` number of losses through games on November 26, 2010
- `week1` number of fumbles on November 6, 2010
- `week2` number of fumbles on November 13, 2010
- `week3` number of fumbles on November 20, 2010

Cases? Variables?

```
data(fumbles, package = "fastR") ## Load data from a pkg
head(fumbles) ## only shows first few obs
```

```
##          team rank W  L week1 week2 week3
## 1  Air Force   53 8  4     4     2     2
## 2       Akron   19 1 11     2     3     2
## 3     Alabama   68 9  3     0     3     2
## 4     Arizona   31 7  4     1     0     2
## 5 Arizona St   94 5  6     2     1     3
## 6    Arkansas   46 9  2     0     1     0
```

Data Frames

- The way I organized the fumbles dataset is called a data frame.
- DBC calls this a "data matrix".
- Each row corresponds to a unique case.
- Each column corresponds to a variable.
- Most datasets in R are `data.frame` objects.

Types of Variables

Variables can be categorical or quantitative.

categorical variable

A categorical variable places each individual into a group or category, such as male or female.

quantitative variable

A quantitative variable has numerical values for which arithmetic operations such as adding and averaging make sense. They measure some characteristic of each case, such as height in centimeters or annual salary in dollars.

label

A label is a special categorical variable used in some datasets to uniquely distinguish the different cases. (e.g. id number, social security number, name)

Categorical or Quantitative?

```
glimpse(trump)
```

```
## Observations: 1,390
## Variables: 6
## $ source   <chr> "Android", "iPhone", "iPhone...
## $ text     <chr> "My economic policy speech w...
## $ hour     <int> 10, 8, 19, 18, 16, 8, 21, 21...
## $ length   <int> 67, 90, 40, 134, 135, 138, 5...
## $ anger    <lgl> FALSE, FALSE, FALSE, TRUE, T...
## $ negative <lgl> FALSE, FALSE, FALSE, TRUE, T...
```

Categorical or Quantitative?

```
head(fumbles)
```

```
##          team rank W  L week1 week2 week3
## 1  Air Force   53 8  4     4     2     2
## 2      Akron   19 1 11     2     3     2
## 3    Alabama   68 9  3     0     3     2
## 4    Arizona   31 7  4     1     0     2
## 5 Arizona St   94 5  6     2     1     3
## 6   Arkansas   46 9  2     0     1     0
```

Categorical or Quantitative?

What if I restructure `fumbles` like this:

```
fumb2 <- gather(fumbles, key = "week",
                value = "fumbles", week1:week3)
head(fumb2)
```

```
##          team rank W  L  week fumbles
## 1  Air Force   53 8  4 week1       4
## 2      Akron   19 1 11 week1       2
## 3    Alabama   68 9  3 week1       0
## 4    Arizona   31 7  4 week1       1
## 5 Arizona St   94 5  6 week1       2
## 6   Arkansas   46 9  2 week1       0
```

Categorical or Quantitative?

- Social security numbers?
- Phone numbers area code?
- I place the class into 10 different groups, labeled 1 through 10. Is group number quantitative or categorical?
- Grade point average (GPA)?