# Data (Finally!)

Understanding Political Numbers

Feb 6, 2019

# Agenda

Technical lesson: The whole lecture

A "vocabulary" of data tables

Variable "coding"

# Setting the scene

# How do you organize data?

# Data tables (a.k.a. "data frames")

| country | continent | year | lifeExp | pop | gdpPercap |
|---|---|---|---|---|---|
| Afghanistan | Asia | 1952 | 28.801 | 8425333 | 779.4453 |
| Afghanistan | Asia | 1957 | 30.332 | 9240934 | 820.8530 |
| Afghanistan | Asia | 1962 | 31.997 | 10267083 | 853.1007 |
| Afghanistan | Asia | 1967 | 34.020 | 11537966 | 836.1971 |
| Afghanistan | Asia | 1972 | 36.088 | 13079460 | 739.9811 |
| Afghanistan | Asia | 1977 | 38.438 | 14880372 | 786.1134 |

(Data from the `gapminder` R package)

# Data tables (a.k.a. "data frames")

| country | continent | year | lifeExp | pop | gdpPercap |
|---|---|---|---|---|---|
| Afghanistan | Asia | 1952 | 28.801 | 8425333 | 779.4453 |
| Afghanistan | Asia | 1957 | 30.332 | 9240934 | 820.8530 |
| Afghanistan | Asia | 1962 | 31.997 | 10267083 | 853.1007 |
| Afghanistan | Asia | 1967 | 34.020 | 11537966 | 836.1971 |
| Afghanistan | Asia | 1972 | 36.088 | 13079460 | 739.9811 |
| Afghanistan | Asia | 1977 | 38.438 | 14880372 | 786.1134 |

(Data from the `gapminder` R package)

Rows are **cases,** a.k.a. "units." What "objects" you studying?

Columns are **variables,** a.k.a. *attributes* or *features* of cases

Cells are **values,** a.k.a. observed measurements of a variable

# Data tables are organized by the "unit of observation"

One row per "unit." What is the unit of observation?

| country | continent | year | lifeExp | pop | gdpPercap |
|---|---|---|---|---|---|
| Afghanistan | Asia | 1952 | 28.801 | 8425333 | 779.4453 |
| Afghanistan | Asia | 1957 | 30.332 | 9240934 | 820.8530 |
| Afghanistan | Asia | 1962 | 31.997 | 10267083 | 853.1007 |
| Afghanistan | Asia | 1967 | 34.020 | 11537966 | 836.1971 |
| Afghanistan | Asia | 1972 | 36.088 | 13079460 | 739.9811 |
| Afghanistan | Asia | 1977 | 38.438 | 14880372 | 786.1134 |

# Data tables are organized by the "unit of observation"

One row per "unit." What is the unit of observation?

| country | continent | year | lifeExp | pop | gdpPercap |
|---|---|---|---|---|---|
| Afghanistan | Asia | 1952 | 28.801 | 8425333 | 779.4453 |
| Afghanistan | Asia | 1957 | 30.332 | 9240934 | 820.8530 |
| Afghanistan | Asia | 1962 | 31.997 | 10267083 | 853.1007 |
| Afghanistan | Asia | 1967 | 34.020 | 11537966 | 836.1971 |
| Afghanistan | Asia | 1972 | 36.088 | 13079460 | 739.9811 |
| Afghanistan | Asia | 1977 | 38.438 | 14880372 | 786.1134 |

**Unit of observation:** what variable (or combination of variables) *uniquely identifies* cases from one another?

# Variables live in columns

Isolating one variable

| continent |
|---|
| Asia |
| Asia |
| Asia |
| Asia |
| Asia |
| Asia |

# Variables live in columns

Isolating one variable

| continent |
| --- |
| Asia |
| Asia |
| Asia |
| Asia |
| Asia |
| Asia |

**Levels**: all possible *values* that a variable could have

```
unique(gapminder$continent)
```

```
## [1] "Asia"     "Europe"   "Africa"   "Americas" "Oceania"
```

# On so many levels...

# On so many levels...

**Levels of a variable** are possible values taken by a variable

- "We expose our plants to various levels of sunlight"

# On so many levels...

**Levels of a variable** are possible values taken by a variable

- "We expose our plants to various levels of sunlight"

**Levels of measurement** describe information contained in measures

- Nominal, ordinal, interval, ratio

# On so many levels...

**Levels of a variable** are possible values taken by a variable

- "We expose our plants to various levels of sunlight"

**Levels of measurement** describe information contained in measures

- Nominal, ordinal, interval, ratio

**Level of analysis** means "unit of analysis"

- Individual-level analysis
- State-level analysis
- Cross-national analysis

# Benefits of tabular data

**Calculations.** If $GDP\ per\ capita = \frac{GDP}{Pop}$, find GDP.

| country | pop | gdpPercap |
|---|---|---|
| Afghanistan | 31889923 | 974.5803 |
| Albania | 3600523 | 5937.0295 |
| Algeria | 33333216 | 6223.3675 |
| Angola | 12420476 | 4797.2313 |
| Argentina | 40301927 | 12779.3796 |
| Australia | 20434176 | 34435.3674 |

# Benefits of tabular data

**Calculations.** If $GDP\ per\ capita = \frac{GDP}{Pop}$, find GDP.

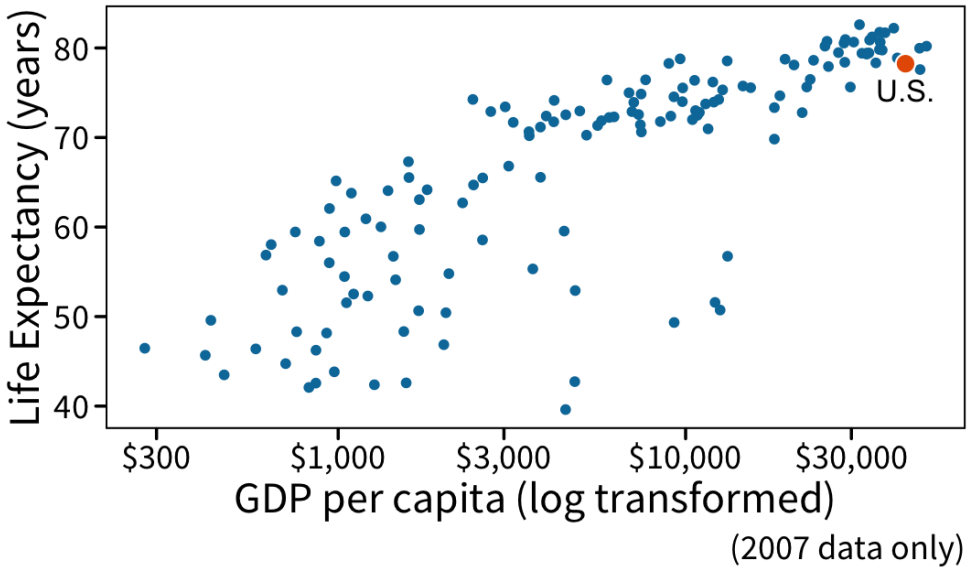
| country | pop | gdpPercap | gdp |
|---|---|---|---|
| Afghanistan | 31889923 | 974.5803 | 31079291949 |
| Albania | 3600523 | 5937.0295 | 21376411360 |
| Algeria | 33333216 | 6223.3675 | 207444851958 |
| Angola | 12420476 | 4797.2313 | 59583895818 |
| Argentina | 40301927 | 12779.3796 | 515033625357 |
| Australia | 20434176 | 34435.3674 | 703658358894 |

# Benefits of tabular data

**Calculations.** If $GDP\ per\ capita = \frac{GDP}{Pop}$, find GDP.

| country | pop | gdpPercap | gdp |
|---|---|---|---|
| Afghanistan | 31889923 | 974.5803 | 31079291949 |
| Albania | 3600523 | 5937.0295 | 21376411360 |
| Algeria | 33333216 | 6223.3675 | 207444851958 |
| Angola | 12420476 | 4797.2313 | 59583895818 |
| Argentina | 40301927 | 12779.3796 | 515033625357 |
| Australia | 20434176 | 34435.3674 | 703658358894 |

**Plotting.** For each unit, variable values serve as $(x, y)$ coordinates.



# Data in the computer

# Data in R

```
# attach the 'gapminder' package
library("gapminder")

# ...which contains this dataset
gapminder
```

Data can be numeric (`1952`) or text (`"Afghanistan"`)

```
## # A tibble: 1,704 x 6
##    country     continent  year lifeExp      pop gdpPercap
##    <chr>       <chr>     <int>   <dbl>    <int>     <dbl>
##  1 Afghanistan Asia       1952    28.8  8425333      779.
##  2 Afghanistan Asia       1957    30.3  9240934      821.
##  3 Afghanistan Asia       1962    32.0 10267083      853.
##  4 Afghanistan Asia       1967    34.0 11537966      836.
##  5 Afghanistan Asia       1972    36.1 13079460      740.
##  6 Afghanistan Asia       1977    38.4 14880372      786.
##  7 Afghanistan Asia       1982    39.9 12881816      978.
##  8 Afghanistan Asia       1987    40.8 13867957      852.
##  9 Afghanistan Asia       1992    41.7 16317921      649.
## 10 Afghanistan Asia       1997    41.8 22227415      635.
## # … with 1,694 more rows
```

# "Coding" and "recoding"

You may encounter categorical data with numeric codes.

| country | continent | continent (numeric) |
|---|---|---|
| Congo, Dem. Rep. | Africa | 1 |
| Panama | Americas | 2 |
| Vietnam | Asia | 3 |
| Finland | Europe | 4 |
| Australia | Oceania | 5 |

# "Coding" and "recoding"

You may encounter categorical data with numeric codes.

| country | continent | continent (numeric) |
|---|---|---|
| Congo, Dem. Rep. | Africa | 1 |
| Panama | Americas | 2 |
| Vietnam | Asia | 3 |
| Finland | Europe | 4 |
| Australia | Oceania | 5 |

Consult the "codebook"

Variable Descriptions:

1) `election`: Election cycle preceded by two-letter state code. Federal candidates have 'fd' as the state code.

2) `cycle`: Four digit number that indicates the two-year election cycle during which the contribution was recorded.

3) `fecyear`: Year listed by the FEC indicating the year of campaign's the target election. The 'election' variable indicates the election cycle during which the contribution was received. But the election can occur in a future cycle—as is the case for senators that fundraise during their first four years in office.

# Practical advice

# Practical advice

When "tidying up" your data

When recording "raw" data

$$Observed = Truth + Bias + Error$$

# What's next?

# What's next?

In section: getting oriented in R

Check out online practice resources

Next week: graphics (in theory) and graphics (in R)

Short Essay 1 due one week from today