LECTURE 15: TIDY DATA

STAT 598z: Introduction to computing for statistics

Vinayak Rao

Department of Statistics, Purdue University

March 9, 2017

Wickham, *Tidy data*, J. of Stat. Software, 2014

- 80% of data analysis: cleaning and preparing data (Dasu & Johnson, 03)
- Often must be repeated over and over again

Wickham, *Tidy data*, J. of Stat. Software, 2014

- 80% of data analysis: cleaning and preparing data (Dasu & Johnson, 03)
- Often must be repeated over and over again

Today, we'll look at the idea of "tidy" data

- A standard way to organize data into tables
- Structure datasets with consistent semantics
- Allows developing tools with tidy inputs and outputs

Wickham, *Tidy data*, J. of Stat. Software, 2014

- 80% of data analysis: cleaning and preparing data (Dasu & Johnson, 03)
- Often must be repeated over and over again

Today, we'll look at the idea of "tidy" data

- A standard way to organize data into tables
- Structure datasets with consistent semantics
- Allows developing tools with tidy inputs and outputs

Note:

- The "best" way of organizing data depends on application
- We're striving to be application independent to allow reuse

Variables and Observations

A dataset is a collection of values (numbers or strings)

A value belongs to a variable and an observation

Variables and Observations

A dataset is a collection of values (numbers or strings)

A value belongs to a variable and an observation

variable an attribute measured across units

`name, height, number of arrests`

observation a collection of measured attributes for a unit

Variables and Observations

A dataset is a collection of values (numbers or strings)

A value belongs to a variable and an observation

variable an attribute measured across units

`name, height, number of arrests`

observation a collection of measured attributes for a unit

Tidy data:

- Every column is a variable
- Every row is an observation

Variables and Observations

Multiple ways of storing same information e.g. rows vs columns

`USJudgeRatings` (reduced)

Columns: Integrity, Demeanor and Diligence (variables)

```
                 INTG DMNR DILG
AARONSON,L.H.     7.9  7.7  7.3
ALEXANDER,J.M.    8.9  8.8  8.5
ARMENTANO,A.J.    8.1  7.8  7.8
```

Variables and Observations

Multiple ways of storing same information e.g. rows vs columns

`USJudgeRatings` (reduced)

Columns: Integrity, Demeanor and Diligence (variables)

```
                INTG DMNR DILG
AARONSON,L.H.    7.9  7.7  7.3
ALEXANDER,J.M.   8.9  8.8  8.5
ARMENTANO,A.J.   8.1  7.8  7.8
```

Untidy/Messy:

```
             AARONSON,L.H. ALEXANDER,J.M. ARMENTANO,A.J.
INTG                   7.9            8.9            8.1
DMNR                   7.7            8.8            7.8
DILG                   7.3            8.5            7.8
```

Variables and Observations

```
                 INTG DMNR DILG
AARONSON,L.H.    7.9  7.7  7.3
ALEXANDER,J.M.   8.9  8.8  8.5
ARMENTANO,A.J.   8.1  7.8  7.8
```

But are the names also values?

Variables and Observations

```
                INTG DMNR DILG
AARONSON,L.H.    7.9  7.7  7.3
ALEXANDER,J.M.   8.9  8.8  8.5
ARMENTANO,A.J.   8.1  7.8  7.8
```

But are the names also values?

```
        NAME     INTG DMNR DILG
1  AARONSON,L.H.  7.9  7.7  7.3
2 ALEXANDER,J.M.  8.9  8.8  8.5
3 ARMENTANO,A.J.  8.1  7.8  7.8
```

Variables and Observations

```
                INTG DMNR DILG
AARONSON,L.H.    7.9  7.7  7.3
ALEXANDER,J.M.   8.9  8.8  8.5
ARMENTANO,A.J.   8.1  7.8  7.8
```

But are the names also values?

```
       NAME      INTG DMNR DILG
1  AARONSON,L.H.  7.9  7.7  7.3
2 ALEXANDER,J.M.  8.9  8.8  8.5
3 ARMENTANO,A.J.  8.1  7.8  7.8
```

What if there’s a new `‘Alexander,J.M.’`?

Variables and Observations

What if there's a new `'Alexander,J.M.'`?

Add a new column '`ID`', and then add row

```
    ID    NAME         INTG DMNR DILG
1   1   AARONSON,L.H.   7.9  7.7  7.3
2   2  ALEXANDER,J.M.   8.9  8.8  8.5
3   3  ARMENTANO,A.J.   8.1  7.8  7.8
4   4  ALEXANDER,J.M.   7.0  8.9  8.3
```

Variables and Observations

We can also turn column names into measured values:

```
   ID    NAME         ATTR ATTR_VAL
1   1  AARONSON,L.H. INTG   7.9
2   2 ALEXANDER,J.M. INTG   8.9
3   3 ARMENTANO,A.J. INTG   8.1
4   4 ALEXANDER,J.M. INTG   7.0
5   1  AARONSON,L.H. DMNR   7.7
6   2 ALEXANDER,J.M. DMNR   8.8
7   3 ARMENTANO,A.J. DMNR   7.8
8   4 ALEXANDER,J.M. DMNR   8.9
9   1  AARONSON,L.H. DILG   7.3
10  2 ALEXANDER,J.M. DILG   8.5
11  3 ARMENTANO,A.J. DILG   7.8
12  4 ALEXANDER,J.M. DILG   8.3
```

Variables and Observations

How, for the last two tables, do we add a new attribute e.g. `RARE` for `ID: 1, AARONSON,L,H`?

Variables and Observations

How, for the last two tables, do we add a new attribute e.g. `RARE` for `ID: 1, AARONSON,L,H`?

In first case, add a new column

```
    ID    NAME         INTG DMNR DILG  RARE
1   1   AARONSON,L.H.  7.9  7.7  7.3  Something
2   2 ALEXANDER,J.M.   8.9  8.8  8.5   NA
3   3 ARMENTANO,A.J.   8.1  7.8  7.8   NA
4   4 ALEXANDER,J.M.   7.0  8.9  8.3   NA
```

In second case, add a new row

Variables and Observations

How, for the last two tables, do we add a new attribute e.g. `RARE` for `ID: 1, AARONSON,L,H`?

In first case, add a new column

```
   ID    NAME        INTG DMNR DILG  RARE
1  1  AARONSON,L.H.  7.9  7.7  7.3  Something
2  2 ALEXANDER,J.M.  8.9  8.8  8.5   NA
3  3 ARMENTANO,A.J.  8.1  7.8  7.8   NA
4  4 ALEXANDER,J.M.  7.0  8.9  8.3   NA
```

In second case, add a new row

The second allows one to ignore structurally missing values (e.g. pregnant males, or temperature on the 31st of February)

Normalization

The tall table from two slides ago also has some redundancy

Don't want the same information `ID` $\sim$ `NAME` in multiple places

Normalization

The tall table from two slides ago also has some redundancy

Don’t want the same information `ID` ~ `NAME` in multiple places

Might help splitting into two tables:

```
   ID  NAME
1  1  AARONSON,L.H.
2  2 ALEXANDER,J.M.
3  3 ARMENTANO,A.J.
4  4 ALEXANDER,J.M.
```

```
    ID  ATTR ATTR_VAL
 1   1  INTG   7.9
 2   2  INTG   8.9
 3   3  INTG   8.1
 4   4  INTG   7.0
 5   1  DMNR   7.7
 ...
```

Tidy data

Tidy data:

- Every column is a variable
- Every row is an observation
- Each type of observational unit forms a table

Tidy data

Tidy data:

- Every column is a variable
- Every row is an observation
- Each type of observational unit forms a table

Messy data:

- Column headers are values, not variable names.
- Multiple variables are stored in one column.
- Variables are stored in both rows and columns.
- Multiple types of observational units stored in same table.
- A single observational unit stored in multiple tables.

The tidyverse package

The “correct” form of your data depends on the application

Worthwhile thinking about this before starting data analysis

The tidyverse package

The “correct” form of your data depends on the application

Worthwhile thinking about this before starting data analysis

Two extremes:

Tall two columns, `variable` and `value`
Wide lots of columns

The tidyverse package

The “correct” form of your data depends on the application

Worthwhile thinking about this before starting data analysis

Two extremes:

Tall	two columns, `variable` and `value`
Wide	lots of columns

Tidy data is often close to tall

Very often the is the most convenient layout

E.g. `ggplot`

Recall the work we needed for a faceted plot in HW3

The tidyverse package

The “correct” form of your data depends on the application

Worthwhile thinking about this before starting data analysis

Two extremes:

Tall two columns, `variable` and `value`
Wide lots of columns

Tidy data is often close to tall

Very often the is the most convenient layout

E.g. `ggplot`

Recall the work we needed for a faceted plot in HW3

`tidyverse` provides convenient tools to shift data between different forms: `gather` and `separate`