

2 - Reshaping

Rearranging Data

Iowa State University

Outline

- ▶ Reshaping Data Using Spreadsheets
- ▶ reshapeGUI
- ▶ melt and cast in the command line

Mac users should be using the terminal server for this session (or should use reshape2 directly).

Reshaping Data

- What do we want to do?

| | Year/Month | Date.1 | Value.1 | Date.2 | Value.2 | Date.3 | Value.3 | Date.4 | Value.4 | Date.5 | Value.5 |
|---|------------|--------|---------|--------|---------|--------|---------|--------|---------|--------|---------|
| 1 | 1994-Nov | | NA | | NA | | NA | 28-Nov | 1.122 | | NA |
| 2 | 1994-Dec | 5-Dec | 1.086 | 12-Dec | 1.057 | 19-Dec | 1.039 | 26-Dec | 1.027 | | NA |
| 3 | 1995-Jan | 2-Jan | 1.025 | 9-Jan | 1.046 | 16-Jan | 1.031 | 23-Jan | 1.054 | 30-Jan | 1.055 |
| 4 | 1995-Feb | 6-Feb | 1.045 | 13-Feb | 1.040 | 20-Feb | 1.031 | 27-Feb | 1.052 | | NA |
| 5 | 1995-Mar | 6-Mar | 1.053 | 13-Mar | 1.042 | 20-Mar | 1.048 | 27-Mar | 1.065 | | NA |
| 6 | 1995-Apr | 3-Apr | 1.091 | 10-Apr | 1.109 | 17-Apr | 1.123 | 24-Apr | 1.148 | | NA |

We have five blocks of weekly dates and gas prices next to each other

Reshaping Data

- What do we want to do?

| | Year/Month | Date.1 | Value.1 | Date.2 | Value.2 | Date.3 | Value.3 | Date.4 | Value.4 | Date.5 | Value.5 |
|---|------------|--------|---------|--------|---------|--------|---------|--------|---------|--------|---------|
| 1 | 1994-Nov | | NA | | NA | | NA | 28-Nov | 1.122 | | NA |
| 2 | 1994-Dec | 5-Dec | 1.086 | 12-Dec | 1.057 | 19-Dec | 1.039 | 26-Dec | 1.027 | | NA |
| 3 | 1995-Jan | 2-Jan | 1.025 | 9-Jan | 1.046 | 16-Jan | 1.031 | 23-Jan | 1.054 | 30-Jan | 1.055 |
| 4 | 1995-Feb | 6-Feb | 1.045 | 13-Feb | 1.040 | 20-Feb | 1.031 | 27-Feb | 1.052 | | NA |
| 5 | 1995-Mar | 6-Mar | 1.053 | 13-Mar | 1.042 | 20-Mar | 1.048 | 27-Mar | 1.065 | | NA |
| 6 | 1995-Apr | 3-Apr | 1.091 | 10-Apr | 1.109 | 17-Apr | 1.123 | 24-Apr | 1.148 | | NA |

We have five blocks of weekly dates and gas prices next to each other

Reshaping Data

- What do we want to do?

| | Year/Month | Date.1 | Value.1 | Date.2 | Value.2 | Date.3 | Value.3 | Date.4 | Value.4 | Date.5 | Value.5 |
|---|------------|--------|---------|--------|---------|--------|---------|--------|---------|--------|---------|
| 1 | 1994-Nov | | NA | | NA | | NA | 28-Nov | 1.122 | | NA |
| 2 | 1994-Dec | 5-Dec | 1.086 | 12-Dec | 1.057 | 19-Dec | 1.039 | 26-Dec | 1.027 | | NA |
| 3 | 1995-Jan | 2-Jan | 1.025 | 9-Jan | 1.046 | 16-Jan | 1.031 | 23-Jan | 1.054 | 30-Jan | 1.055 |
| 4 | 1995-Feb | 6-Feb | 1.045 | 13-Feb | 1.040 | 20-Feb | 1.031 | 27-Feb | 1.052 | | NA |
| 5 | 1995-Mar | 6-Mar | 1.053 | 13-Mar | 1.042 | 20-Mar | 1.048 | 27-Mar | 1.065 | | NA |
| 6 | 1995-Apr | 3-Apr | 1.091 | 10-Apr | 1.109 | 17-Apr | 1.123 | 24-Apr | 1.148 | | NA |

We have five blocks of weekly dates and gas prices next to each other

Reshaping Data

- What do we want to do?

| | Year/Month | Date.1 | Value.1 | Date.2 | Value.2 | Date.3 | Value.3 | Date.4 | Value.4 | Date.5 | Value.5 |
|---|------------|--------|---------|--------|---------|--------|---------|--------|---------|--------|---------|
| 1 | 1994-Nov | | NA | | NA | | NA | 28-Nov | 1.122 | | NA |
| 2 | 1994-Dec | 5-Dec | 1.086 | 12-Dec | 1.057 | 19-Dec | 1.039 | 26-Dec | 1.027 | | NA |
| 3 | 1995-Jan | 2-Jan | 1.025 | 9-Jan | 1.046 | 16-Jan | 1.031 | 23-Jan | 1.054 | 30-Jan | 1.055 |
| 4 | 1995-Feb | 6-Feb | 1.045 | 13-Feb | 1.040 | 20-Feb | 1.031 | 27-Feb | 1.052 | | NA |
| 5 | 1995-Mar | 6-Mar | 1.053 | 13-Mar | 1.042 | 20-Mar | 1.048 | 27-Mar | 1.065 | | NA |
| 6 | 1995-Apr | 3-Apr | 1.091 | 10-Apr | 1.109 | 17-Apr | 1.123 | 24-Apr | 1.148 | | NA |

We have five blocks of weekly dates and gas prices next to each other

Reshaping Data

- What do we want to do?

| | Year/Month | Date.1 | Value.1 | Date.2 | Value.2 | Date.3 | Value.3 | Date.4 | Value.4 | Date.5 | Value.5 |
|---|------------|--------|---------|--------|---------|--------|---------|--------|---------|--------|---------|
| 1 | 1994-Nov | | NA | | NA | | NA | 28-Nov | 1.122 | | NA |
| 2 | 1994-Dec | 5-Dec | 1.086 | 12-Dec | 1.057 | 19-Dec | 1.039 | 26-Dec | 1.027 | | NA |
| 3 | 1995-Jan | 2-Jan | 1.025 | 9-Jan | 1.046 | 16-Jan | 1.031 | 23-Jan | 1.054 | 30-Jan | 1.055 |
| 4 | 1995-Feb | 6-Feb | 1.045 | 13-Feb | 1.040 | 20-Feb | 1.031 | 27-Feb | 1.052 | | NA |
| 5 | 1995-Mar | 6-Mar | 1.053 | 13-Mar | 1.042 | 20-Mar | 1.048 | 27-Mar | 1.065 | | NA |
| 6 | 1995-Apr | 3-Apr | 1.091 | 10-Apr | 1.109 | 17-Apr | 1.123 | 24-Apr | 1.148 | | NA |

We have five blocks of weekly dates and gas prices next to each other

Reshaping Data

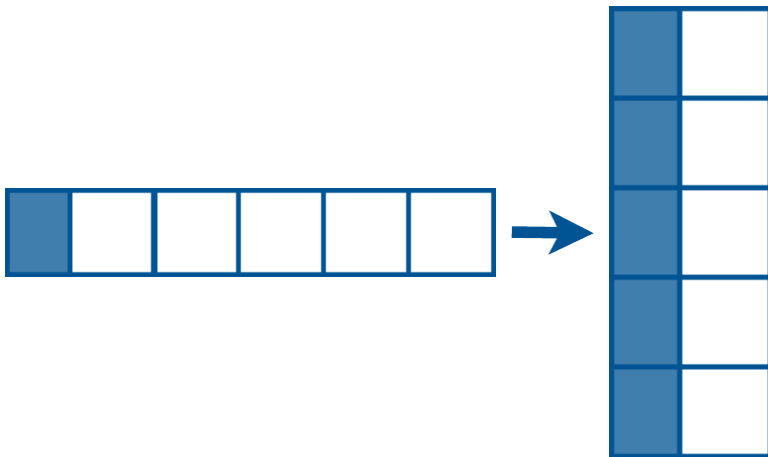
- What do we want to do?

| | Year/Month | Date.1 | Value.1 | Date.2 | Value.2 | Date.3 | Value.3 | Date.4 | Value.4 | Date.5 | Value.5 |
|---|------------|--------|---------|--------|---------|--------|---------|--------|---------|--------|---------|
| 1 | 1994-Nov | | NA | | NA | | NA | 28-Nov | 1.122 | | NA |
| 2 | 1994-Dec | 5-Dec | 1.086 | 12-Dec | 1.057 | 19-Dec | 1.039 | 26-Dec | 1.027 | | NA |
| 3 | 1995-Jan | 2-Jan | 1.025 | 9-Jan | 1.046 | 16-Jan | 1.031 | 23-Jan | 1.054 | 30-Jan | 1.055 |
| 4 | 1995-Feb | 6-Feb | 1.045 | 13-Feb | 1.040 | 20-Feb | 1.031 | 27-Feb | 1.052 | | NA |
| 5 | 1995-Mar | 6-Mar | 1.053 | 13-Mar | 1.042 | 20-Mar | 1.048 | 27-Mar | 1.065 | | NA |
| 6 | 1995-Apr | 3-Apr | 1.091 | 10-Apr | 1.109 | 17-Apr | 1.123 | 24-Apr | 1.148 | | NA |

We have five blocks of weekly dates and gas prices next to each other

Reshaping Data

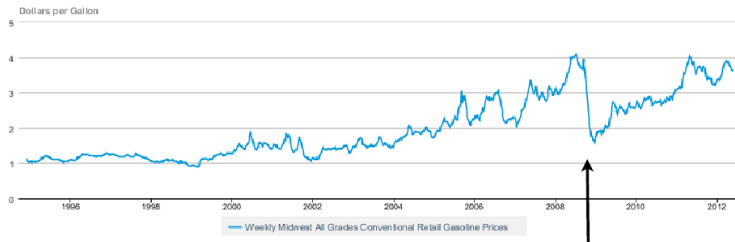
- What do we want to do?



Reshaping Data

- Earlier we read the midwest gas prices

Weekly Midwest All Grades Conventional Retail Gasoline Prices



 Source: U.S. Energy Information Administration

Oct 2008

Your Turn

- ▶ Use a spreadsheet program to reshape the Midwest Gas Price data from “wide” form to “long” form

The reshape GUI

The screenshot shows the 'Data Reshaping & Aggregating' window. On the left, the 'Data Selection' panel lists 'Current_Data_Frames_in_Workspace' with items: gp, gp_data (selected), gp_names, gp2, and mainland. Below this is a 'Refresh Workspace' button. To the right, the 'File Loading/Exporting' section has 'Load .csv File' with 'Object Name: newData' and an unchecked 'Header?' checkbox. The 'Export to .csv File' section has a 'File Name:' field and an unchecked 'Include Row Names?' checkbox. Below these are 'Data Manipulation' buttons: 'melt', 'cast', 'ddply', and a help icon '?'. The bottom section, 'Data Preview: gp_data (211 rows, 11 columns)', displays a table with 16 rows and 11 columns.

| Row.names | Year/Month | Date.1 | Value.1 | Date.2 | Value.2 | Date.3 | Value.3 | Date.4 | Value.4 | Date.5 | Value.5 |
|-----------|------------|--------|---------|--------|---------|--------|---------|--------|---------|--------|---------|
| 1 | 1994-Nov | | NA | | NA | | NA | 28-Nov | 1.122 | | NA |
| 2 | 1994-Dec | 5-Dec | 1.086 | 12-Dec | 1.057 | 19-Dec | 1.039 | 26-Dec | 1.027 | | NA |
| 3 | 1995-Jan | 2-Jan | 1.025 | 9-Jan | 1.046 | 16-Jan | 1.031 | 23-Jan | 1.054 | 30-Jan | 1.055 |
| 4 | 1995-Feb | 6-Feb | 1.045 | 13-Feb | 1.04 | 20-Feb | 1.031 | 27-Feb | 1.052 | | NA |
| 5 | 1995-Mar | 6-Mar | 1.053 | 13-Mar | 1.042 | 20-Mar | 1.048 | 27-Mar | 1.065 | | NA |
| 6 | 1995-Apr | 3-Apr | 1.091 | 10-Apr | 1.109 | 17-Apr | 1.123 | 24-Apr | 1.148 | | NA |
| 7 | 1995-May | 1-May | 1.144 | 8-May | 1.177 | 15-May | 1.183 | 22-May | 1.21 | 29-May | 1.202 |
| 8 | 1995-Jun | 5-Jun | 1.202 | 12-Jun | 1.195 | 19-Jun | 1.181 | 26-Jun | 1.169 | | NA |
| 9 | 1995-Jul | 3-Jul | 1.158 | 10-Jul | 1.141 | 17-Jul | 1.125 | 24-Jul | 1.112 | 31-Jul | 1.097 |
| 10 | 1995-Aug | 7-Aug | 1.102 | 14-Aug | 1.102 | 21-Aug | 1.108 | 28-Aug | 1.096 | | NA |
| 11 | 1995-Sep | 4-Sep | 1.098 | 11-Sep | 1.094 | 18-Sep | 1.097 | 25-Sep | 1.098 | | NA |
| 12 | 1995-Oct | 2-Oct | 1.091 | 9-Oct | 1.076 | 16-Oct | 1.059 | 23-Oct | 1.05 | 30-Oct | 1.033 |
| 13 | 1995-Nov | 6-Nov | 1.04 | 13-Nov | 1.046 | 20-Nov | 1.042 | 27-Nov | 1.047 | | NA |
| 14 | 1995-Dec | 4-Dec | 1.053 | 11-Dec | 1.049 | 18-Dec | 1.089 | 25-Dec | 1.087 | | NA |
| 15 | 1996-Jan | 1-Jan | 1.089 | 8-Jan | 1.108 | 15-Jan | 1.106 | 22-Jan | 1.092 | 29-Jan | 1.079 |
| 16 | 1996-Feb | 5-Feb | 1.079 | 12-Feb | 1.078 | 19-Feb | 1.089 | 26-Feb | 1.131 | | NA |

The reshape GUI

Data Reshaping & Aggregating

Data Selection | melt: gp_data | melt: gp_data

| ID_Variables | Unused_Variables | Measure_Variables |
|--------------|--|---|
| Year/Month | Date.1 Date.2 Date.3 Date.4 Date.5 | Value.1 Value.2 Value.3 Value.4 Value.5 |

Reset

Code: `gp_data.melt <- melt(data = gp_data, id.vars=c('Year/Month'), measure.vars=c('Value.1','Value.2','Value.3','Value.4','Value.5'))` Preview Execute

Raw Data

| Row.names | Year/Month | Date.1 | Value.1 | Date.2 | Value.2 |
|-----------|------------|--------|---------|--------|---------|
| 1 | 1994-Nov | | NA | | NA |
| 2 | 1994-Dec | 5-Dec | 1.086 | 12-Dec | 1.057 |
| 3 | 1995-Jan | 2-Jan | 1.025 | 9-Jan | 1.046 |
| 4 | 1995-Feb | 6-Feb | 1.045 | 13-Feb | 1.04 |
| 5 | 1995-Mar | 6-Mar | 1.053 | 13-Mar | 1.042 |
| 6 | 1995-Apr | 3-Apr | 1.091 | 10-Apr | 1.109 |
| 7 | 1995-May | 1-May | 1.144 | 8-May | 1.177 |
| 8 | 1995-Jun | 5-Jun | 1.202 | 12-Jun | 1.195 |
| 9 | 1995-Jul | 3-Jul | 1.158 | 10-Jul | 1.141 |
| 10 | 1995-Aug | 7-Aug | 1.102 | 14-Aug | 1.102 |
| 11 | 1995-Sep | 4-Sep | 1.098 | 11-Sep | 1.094 |

Melted Data

| Row.names | Year/Month | variable | value |
|-----------|------------|----------|-------|
| 1 | 1994-Nov | Value.1 | NA |
| 2 | 1994-Dec | Value.1 | 1.086 |
| 3 | 1995-Jan | Value.1 | 1.025 |
| 4 | 1995-Feb | Value.1 | 1.045 |
| 5 | 1995-Mar | Value.1 | 1.053 |
| ... | | | |
| 1050 | 2011-Dec | Value.5 | NA |
| 1051 | 2012-Jan | Value.5 | 3.377 |
| 1052 | 2012-Feb | Value.5 | NA |
| 1053 | 2012-Mar | Value.5 | NA |
| 1054 | 2012-Apr | Value.5 | 3.752 |

Melting Gas Prices

```
library(reshape2)
gp_data.melt <- melt(data=gp_data, id.vars="YM",
  measure.vars=c("Value.1", "Value.2", "Value.3", "Value.4", "Value.5"))
gp_prices <- gp_data.melt
head(gp_prices)
```

| ## | | YM | variable | value |
|------|----------|---------|----------|-------|
| ## 1 | 1994-Nov | Value.1 | NA | |
| ## 2 | 1994-Dec | Value.1 | 1.086 | |
| ## 3 | 1995-Jan | Value.1 | 1.025 | |
| ## 4 | 1995-Feb | Value.1 | 1.045 | |
| ## 5 | 1995-Mar | Value.1 | 1.053 | |
| ## 6 | 1995-Apr | Value.1 | 1.091 | |

Your Turn

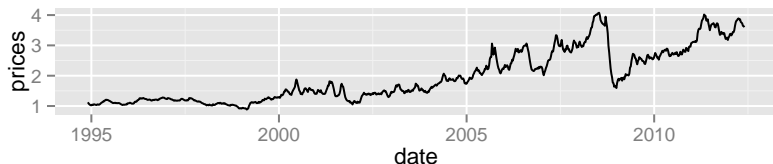
- ▶ Open the reshapeGUI
- ▶ load `gp_data`
- ▶ use `melt` to get one column of dates, similar to how we got a single column of weekly gas prices
- ▶ export the data as '`gp_dates`'

Piecing datasets together

```
gasprices <- gp_dates
gasprices$prices <- gp_prices$value

#--
tmp <- with(gasprices, paste(YM, value, sep="/"))
gasprices$date <- as.Date(tmp, format="%Y-%b/%d-%b")

#--
qplot(date, prices, data=gasprices, geom="line")
```



A Closer Look at reshape

First, melt

- ▶ First we need to melt the data into a long form
- ▶ This form is useful for “casting” it into new formats
- ▶ When melting, you need to specify the **measured** variables and the **identifiers**

```
melt(data, measure.var=..., id.var=...)
```

Measured variables & identifiers

Identifiers/Keys:

- ▶ Identify a record (must be unique)
- ▶ Example: Indices on an random variable
- ▶ Fixed by design of experiment (known in advance)
- ▶ May be single or composite (may have one or more variables)

Measured Variables:

- ▶ Collected during the experiment (not known in advance)
- ▶ Usually numeric quantities

Example: French Fries

During a ten week sensory experiment, 12 individuals were asked to assess taste of french fries on several scales (how potato-y, buttery, grassy, rancid, paint-y do the fries taste?)

French fries were fried in one of three different oils, and each week individuals had to assess six batches of french fries (all three oils, replicated twice)

What are the identifiers?

Example: French Fries

During a **ten week sensory experiment**, 12 **individuals** were asked to assess taste of french fries on several scales (how potato-y, buttery, grassy, rancid, paint-y do the fries taste?)

French fries were fried in one of three **different oils**, and each week individuals had to assess six batches of french fries (all three oils, **replicated** twice)

Example: French Fries

```
library("reshape2")
head(french_fries)
##      time treatment subject rep potato buttery grassy rancid painty
## 61      1          1       3   1    2.9      0.0   0.0   0.0   5.5
## 25      1          1       3   2   14.0      0.0   0.0   1.1   0.0
## 62      1          1      10   1   11.0      6.4   0.0   0.0   0.0
## 26      1          1      10   2    9.9      5.9   2.9   2.2   0.0
## 63      1          1      15   1    1.2      0.1   0.0   1.1   5.1
## 27      1          1      15   2    8.8      3.0   3.6   1.5   2.3
```

```
ffm <- melt(french_fries, id.vars=1:4)
head(ffm)
##      time treatment subject rep variable value
## 1      1          1       3   1  potato    2.9
## 2      1          1       3   2  potato   14.0
## 3      1          1      10   1  potato   11.0
## 4      1          1      10   2  potato    9.9
## 5      1          1      15   1  potato    1.2
## 6      1          1      15   2  potato    8.8
```

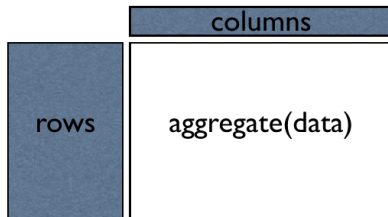
```
summary(ffm)
##      time      treatment      subject      rep      variable      value
## 1      : 360    1:1160    10 : 300    Min.   :1.0    potato :696    Min.   : 0.00
## 2      : 360    2:1160    15 : 300    1st Qu.:1.0    buttery:696    1st Qu.: 0.00
## 3      : 360    3:1160    16 : 300    Median :1.5    grassy :696    Median : 1.50
## 4      : 360                19 : 300    Mean   :1.5    rancid :696    Mean   : 3.16
## 5      : 360                51 : 300    3rd Qu.:2.0    painty :696    3rd Qu.: 5.50
## 6      : 360                52 : 300    Max.   :2.0                Max.   :14.90
## (Other):1320                (Other):1680                NA's   :9
```

Your Turn

- ▶ Explore inter-replicate consistency
- ▶ Pattern of missingness?

Casting

```
cast(molten, rows ~ columns, aggregate)
```



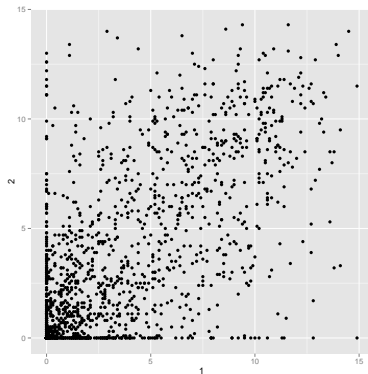
Casting

- ▶ Just like pivot tables and facetting plots
- ▶ Row variables, column variables, and a summary function (sum, mean, max...)

```
cast(molten, row~col, summary)
cast(molten, row1 + row2~col, summary)
cast(molten, row~., summary)
cast(molten, .~col, summary)
```

Inter-rep consistency

```
reps <- dcast(ffm, time+subject+  
  treatment+variable~rep)  
head(reps)  
##   time subject treatment variable    1    2  
## 1     1       3         1  potato  2.9 14.0  
## 2     1       3         1  buttery  0.0  0.0  
## 3     1       3         1   grassy  0.0  0.0  
## 4     1       3         1  rancid   0.0  1.1  
## 5     1       3         1  painty   5.5  0.0  
## 6     1       3         2  potato 13.9 13.4  
  
qplot(`1`, `2`, data=reps)
```



Your Turn

- ▶ How do average ratings by scale (potato-y, buttery, ...) vary over time?

Hint: Start with a cast by scale, then include averages by scale, then include time...

- ▶ Challenge: find the correlation between replicate 1 and replicate 2 over time.