

Tidy data & tidy tools

Hadley Wickham

Assistant Professor / Dobelman Family Junior Chair
Department of Statistics / Rice University

October 2011



1. What is tidy data?

2. Data tidying (3/5)

3. Tidy tools

4. Case study

**What is
tidy data?**

What is tidy data?

- **Data that makes data analysis easy**
- Data that is easy to model, visualise and transform.
- A step along the road to clean data.
- Relational database theory for statisticians

	Pregnant	Not pregnant
Male	0	5
Female	1	4

There are three variables in this data set.
What are they?

pregnant	sex	freq
no	female	4
no	male	5
yes	female	1
yes	male	0

Storage	Meaning
Table / File	Data set
Rows	Observations
Columns	Variables

Data tidying

Causes of messiness

- 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 experimental unit stored in the same table
- One type of experimental unit stored in multiple tables

```
# Tools
```

```
library(reshape2)
```

```
?melt
```

```
?dcast
```

```
library(stringr) # regular expressions
```

```
?str_replace
```

```
?str_sub
```

```
?str_match
```

```
?str_split_fixed
```

```
library(plyr) # optional, but nice
```

```
?arrange
```

Column headers
values, not variable
names

	religion	<\$10k	\$10-20k	\$20-30k	\$30-40k	\$40-50k	\$50-75k
1	Agnostic	27	34	60	81	76	137
2	Atheist	12	27	37	52	35	70
3	Buddhist	27	21	30	34	33	58
4	Catholic	418	617	732	670	638	1116
5	Don't know/refused	15	14	15	11	10	35
6	Evangelical Prot	575	869	1064	982	881	1486
7	Hindu	1	9	7	9	11	34
8	Historically Black Prot	228	244	236	238	197	223
9	Jehovah's Witness	20	27	24	24	21	30
10	Jewish	19	19	25	25	30	95
11	Mainline Prot	289	495	619	655	651	1107
12	Mormon	29	40	48	51	56	112
13	Muslim	6	7	9	10	9	23
14	Orthodox	13	17	23	32	32	47
15	Other Christian	9	7	11	13	13	14
16	Other Faiths	20	33	40	46	49	63
17	Other World Religions	5	2	3	4	2	7
18	Unaffiliated	217	299	374	365	341	528

	religion	<\$10k	\$10-20k	\$20-30k	\$30-40k	\$40-50k	\$50-75k
1	Agnostic	27	34	60	81	76	137
2	Atheist	12	27	37	52	35	70
3	Buddhist	27	21	30	34	33	58
4	Catholic	418	617	732	670	638	1116
5	Don't know/refused	15	14	15	11	10	35
6	Evangelical Prot	575	869	1064	982	881	1486
7	Hindu	1	9	7	9	11	34
8	Historically Black Prot	228	244	236	238	197	223
9	Jehovah's Witness	20	27	24	24	21	30
10	Jewish	19	19	25	25	30	95
11	Mainline Prot	289	495	619	655	651	1107
12	Mormon	29	40	48	51	56	112
13	Muslim	6	7	9	10	9	23
14	Orthodox	13	17	23	32	32	47
15	Other Christian	9	7	11	13	13	14
16	Other Faiths	20	33	40	46	49	63
17	Other World Religions	5	2	3	4	2	7

Un# What are the variables in this dataset?
 # Discuss with your neighbour for 1 minute

```
raw <- read.delim("pew.txt", check.names = F,  
  stringsAsFactors = F)  
  
# Fixing this problem is easy. We use melt, from  
# reshape2, with two arguments, the input data, and  
# the columns which are already variables:  
  
library(reshape2)  
tidy <- melt(raw, "religion")  
  
head(tidy)  
  
# We can now tweak the variable names  
names(tidy) <- c("religion", "income", "n")
```

	religion	income	n		religion	income	n
1	Agnostic	<\$10k	27	26	Historically Black Prot	\$10-20k	244
2	Atheist	<\$10k	12	27	Jehovah's Witness	\$10-20k	27
3	Buddhist	<\$10k	27	28	Jewish	\$10-20k	19
4	Catholic	<\$10k	418	29	Mainline Prot	\$10-20k	495
5	Don't know/refused	<\$10k	15	30	Mormon	\$10-20k	40
6	Evangelical Prot	<\$10k	575	31	Muslim	\$10-20k	7
7	Hindu	<\$10k	1	32	Orthodox	\$10-20k	17
8	Historically Black Prot	<\$10k	228	33	Other Christian	\$10-20k	7
9	Jehovah's Witness	<\$10k	20	34	Other Faiths	\$10-20k	33
10	Jewish	<\$10k	19	35	Other World Religions	\$10-20k	2
11	Mainline Prot	<\$10k	289	36	Unaffiliated	\$10-20k	299
12	Mormon	<\$10k	29	37	Agnostic	\$20-30k	60
13	Muslim	<\$10k	6	38	Atheist	\$20-30k	37
14	Orthodox	<\$10k	13	39	Buddhist	\$20-30k	30
15	Other Christian	<\$10k	9	40	Catholic	\$20-30k	732
16	Other Faiths	<\$10k	20	41	Don't know/refused	\$20-30k	15
17	Other World Religions	<\$10k	5	42	Evangelical Prot	\$20-30k	1064
18	Unaffiliated	<\$10k	217	43	Hindu	\$20-30k	7
19	Agnostic	\$10-20k	34	44	Historically Black Prot	\$20-30k	236
20	Atheist	\$10-20k	27	45	Jehovah's Witness	\$20-30k	24
21	Buddhist	\$10-20k	21	46	Jewish	\$20-30k	25
22	Catholic	\$10-20k	617	47	Mainline Prot	\$20-30k	619
23	Don't know/refused	\$10-20k	14	48	Mormon	\$20-30k	48
24	Evangelical Prot	\$10-20k	869	49	Muslim	\$20-30k	9
25	Hindu	\$10-20k	9	50	Orthodox	\$20-30k	23

Multiple variables in one column

	iso2	year	m04	m514	m014	m1524	m2534	m3544	m4554	m5564	m65	mu	f04	f514	f014
1	AD	1989	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2	AD	1990	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
3	AD	1991	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
4	AD	1992	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
5	AD	1993	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
6	AD	1994	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
7	AD	1996	NA	NA	0	0	0	4	1	0	0	NA	NA	NA	0
8	AD	1997	NA	NA	0	0	1	2	2	1	6	NA	NA	NA	0
9	AD	1998	NA	NA	0	0	0	1	0	0	0	NA	NA	NA	NA
10	AD	1999	NA	NA	0	0	0	1	1	0	0	NA	NA	NA	0
11	AD	2000	NA	NA	0	0	1	0	0	0	0	NA	NA	NA	NA
12	AD	2001	NA	NA	0	NA	NA	2	1	NA	NA	NA	NA	NA	NA
13	AD	2002	NA	NA	0	0	0	1	0	0	0	NA	NA	NA	0
14	AD	2003	NA	NA	0	0	0	1	2	0	0	NA	NA	NA	0
15	AD	2004	NA	NA	0	0	0	1	1	0	0	NA	NA	NA	0
16	AD	2005	0	0	0	0	1	1	0	0	0	0	0	0	0
17	AD	2006	0	0	0	1	1	2	0	1	1	0	0	0	0
18	AD	2007	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
19	AD	2008	0	0	0	0	0	0	1	0	0	0	0	0	0
20	AE	1980	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

	iso2	year	m04	m514	m014	m1524	m2534	m3544	m4554	m5564	m65	mu	f04	f514	f014
1	AD	1989	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2	AD	1990	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
3	AD	1991	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
4	AD	1992	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
5	AD	1993	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
6	AD	1994	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
7	AD	1996	NA	NA	0	0	0	4	1	0	0	NA	NA	NA	0
8	AD	1997	NA	NA	0	0	1	2	2	1	6	NA	NA	NA	0
9	AD	1998	NA	NA	0	0	0	1	0	0	0	NA	NA	NA	NA
10	AD	1999	NA	NA	0	0	0	1	1	0	0	NA	NA	NA	0
11	AD	2000	NA	NA	0	0	1	0	0	0	0	NA	NA	NA	NA
12	AD	2001	NA	NA	0	NA	NA	2	1	NA	NA	NA	NA	NA	NA
13	AD	2002	NA	NA	0	0	0	1	0	0	0	NA	NA	NA	0
14	AD	2003	NA	NA	0	0	0	1	2	0	0	NA	NA	NA	0
15	AD	2004	NA	NA	0	0	0	1	1	0	0	NA	NA	NA	0
16	AD	2005	0	0	0	0	1	1	0	0	0	0	0	0	0
17	AD	2006	0	0	0	1	1	2	0	1	1	0	0	0	0
18	AD	2007	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

What are the variables in this dataset?
 # Discuss with your neighbour for 1 minute
 # Hint: f = female, u = unknown, 1524 = 15-25

```
raw <- read.csv("tb.csv", stringsAsFactors = FALSE)
raw$new_sp <- NULL

names(raw) <- str_replace(names(raw), "new_sp_", "")

# na.rm = TRUE is useful if the missings don't have
# any meaning
tidy <- melt(raw, id = c("iso2", "year"),
  na.rm = TRUE)
names(tidy)[4] <- "cases"

# Often a good idea to ensure the rows are ordered
# by the variables
tidy <- arrange(tidy, iso2, variable, year)
```

	iso2	year	variable	cases
1	AD	2005	m04	0
2	AD	2006	m04	0
3	AD	2008	m04	0
4	AD	2005	m514	0
5	AD	2006	m514	0
6	AD	2008	m514	0
7	AD	1996	m014	0
8	AD	1997	m014	0
9	AD	1998	m014	0
10	AD	1999	m014	0
11	AD	2000	m014	0
12	AD	2001	m014	0
13	AD	2002	m014	0
14	AD	2003	m014	0
15	AD	2004	m014	0
16	AD	2005	m014	0
17	AD	2006	m014	0
18	AD	2008	m014	0
19	AD	1996	m1524	0
20	AD	1997	m1524	0
21	AD	1998	m1524	0
22	AD	1999	m1524	0
23	AD	2000	m1524	0
24	AD	2002	m1524	0
25	AD	2003	m1524	0

	iso2	year	variable	cases
26	AD	2004	m1524	0
27	AD	2005	m1524	0
28	AD	2006	m1524	1
29	AD	2008	m1524	0
30	AD	1996	m2534	0
31	AD	1997	m2534	1
32	AD	1998	m2534	0
33	AD	1999	m2534	0
34	AD	2000	m2534	1
35	AD	2002	m2534	0
36	AD	2003	m2534	0
37	AD	2004	m2534	0
38	AD	2005	m2534	1
39	AD	2006	m2534	1
40	AD	2008	m2534	0
41	AD	1996	m3544	4
42	AD	1997	m3544	2
43	AD	1998	m3544	1
44	AD	1999	m3544	1
45	AD	2000	m3544	0
46	AD	2001	m3544	2
47	AD	2002	m3544	1
48	AD	2003	m3544	1
49	AD	2004	m3544	1
50	AD	2005	m3544	1

```
str_sub(tidy$variable, 1, 1)
str_sub(tidy$variable, 2)

ages <- c("04" = "0-4", "514" = "5-14",
  "014" = "0-14", "1524" = "15-24", "2534" = "25-34",
  "3544" = "35-44", "4554" = "45-54", "5564" = "55-64",
  "65" = "65+", "u" = NA)
ages[str_sub(tidy$variable, 2)]

tidy$sex <- str_sub(tidy$variable, 1, 1)
tidy$age <- factor(ages[str_sub(tidy$variable, 2)],
  levels = ages)
tidy$variable <- NULL

tidy <- tidy[c("iso2", "year", "sex", "age", "cases")]
```

	iso2	year	sex	age	cases
1	AD	2005	m	0-4	0
2	AD	2006	m	0-4	0
3	AD	2008	m	0-4	0
4	AD	2005	m	5-14	0
5	AD	2006	m	5-14	0
6	AD	2008	m	5-14	0
7	AD	1996	m	0-14	0
8	AD	1997	m	0-14	0
9	AD	1998	m	0-14	0
10	AD	1999	m	0-14	0
11	AD	2000	m	0-14	0
12	AD	2001	m	0-14	0
13	AD	2002	m	0-14	0
14	AD	2003	m	0-14	0
15	AD	2004	m	0-14	0
16	AD	2005	m	0-14	0
17	AD	2006	m	0-14	0
18	AD	2008	m	0-14	0
19	AD	1996	m	15-24	0
20	AD	1997	m	15-24	0
21	AD	1998	m	15-24	0
22	AD	1999	m	15-24	0
23	AD	2000	m	15-24	0
24	AD	2002	m	15-24	0
25	AD	2003	m	15-24	0

	iso2	year	sex	age	cases
26	AD	2004	m	15-24	0
27	AD	2005	m	15-24	0
28	AD	2006	m	15-24	1
29	AD	2008	m	15-24	0
30	AD	1996	m	25-34	0
31	AD	1997	m	25-34	1
32	AD	1998	m	25-34	0
33	AD	1999	m	25-34	0
34	AD	2000	m	25-34	1
35	AD	2002	m	25-34	0
36	AD	2003	m	25-34	0
37	AD	2004	m	25-34	0
38	AD	2005	m	25-34	1
39	AD	2006	m	25-34	1
40	AD	2008	m	25-34	0
41	AD	1996	m	35-44	4
42	AD	1997	m	35-44	2
43	AD	1998	m	35-44	1
44	AD	1999	m	35-44	1
45	AD	2000	m	35-44	0
46	AD	2001	m	35-44	2
47	AD	2002	m	35-44	1
48	AD	2003	m	35-44	1
49	AD	2004	m	35-44	1
50	AD	2005	m	35-44	1

Variables in rows and columns

	id	year	month	element	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	d11	d12
1	MX000017004	2010	1	TMAX	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2	MX000017004	2010	1	TMIN	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
3	MX000017004	2010	2	TMAX	NA	273	241	NA	NA	NA	NA	NA	NA	NA	297	NA
4	MX000017004	2010	2	TMIN	NA	144	144	NA	NA	NA	NA	NA	NA	NA	134	NA
5	MX000017004	2010	3	TMAX	NA	NA	NA	NA	321	NA	NA	NA	NA	345	NA	NA
6	MX000017004	2010	3	TMIN	NA	NA	NA	NA	142	NA	NA	NA	NA	168	NA	NA
7	MX000017004	2010	4	TMAX	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
8	MX000017004	2010	4	TMIN	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
9	MX000017004	2010	5	TMAX	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
10	MX000017004	2010	5	TMIN	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
11	MX000017004	2010	6	TMAX	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
12	MX000017004	2010	6	TMIN	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
13	MX000017004	2010	7	TMAX	NA	NA	286	NA	NA	NA	NA	NA	NA	NA	NA	NA
14	MX000017004	2010	7	TMIN	NA	NA	175	NA	NA	NA	NA	NA	NA	NA	NA	NA
15	MX000017004	2010	8	TMAX	NA	NA	NA	NA	296	NA	NA	290	NA	NA	NA	NA
16	MX000017004	2010	8	TMIN	NA	NA	NA	NA	158	NA	NA	173	NA	NA	NA	NA
17	MX000017004	2010	10	TMAX	NA	NA	NA	NA	270	NA	281	NA	NA	NA	NA	NA
18	MX000017004	2010	10	TMIN	NA	NA	NA	NA	140	NA	129	NA	NA	NA	NA	NA
19	MX000017004	2010	11	TMAX	NA	313	NA	272	263	NA	NA	NA	NA	NA	NA	NA
20	MX000017004	2010	11	TMIN	NA	163	NA	120	79	NA	NA	NA	NA	NA	NA	NA

	id	year	month	element	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	d11	d12
1	MX000017004	2010	1	TMAX	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2	MX000017004	2010	1	TMIN	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
3	MX000017004	2010	2	TMAX	NA	273	241	NA	NA	NA	NA	NA	NA	NA	297	NA
4	MX000017004	2010	2	TMIN	NA	144	144	NA	NA	NA	NA	NA	NA	NA	134	NA
5	MX000017004	2010	3	TMAX	NA	NA	NA	NA	321	NA	NA	NA	NA	345	NA	NA
6	MX000017004	2010	3	TMIN	NA	NA	NA	NA	142	NA	NA	NA	NA	168	NA	NA
7	MX000017004	2010	4	TMAX	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
8	MX000017004	2010	4	TMIN	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
9	MX000017004	2010	5	TMAX	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
10	MX000017004	2010	5	TMIN	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
11	MX000017004	2010	6	TMAX	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
12	MX000017004	2010	6	TMIN	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
13	MX000017004	2010	7	TMAX	NA	NA	286	NA	NA	NA	NA	NA	NA	NA	NA	NA
14	MX000017004	2010	7	TMIN	NA	NA	175	NA	NA	NA	NA	NA	NA	NA	NA	NA
15	MX000017004	2010	8	TMAX	NA	NA	NA	NA	296	NA	NA	290	NA	NA	NA	NA
16	MX000017004	2010	8	TMIN	NA	NA	NA	NA	158	NA	NA	173	NA	NA	NA	NA

What are the variables in this dataset?
Discuss with your neighbour for 1 minute
Hint: TMIN = minimum temperature,
id = weather station identifier

```
raw <- read.table("weather.txt",
  stringsAsFactors = FALSE)

raw1 <- melt(raw, id = 1:4, na.rm = T)
raw1$day <- as.integer(
  str_replace(raw1$variable, "d", ""))
raw1$variable <- NULL
raw1$element <- tolower(raw1$element)

raw1 <- raw1[c("id", "year", "month", "day",
  "element", "value")]
raw1 <- arrange(raw1, year, month, day, element)
```

	id	year	month	day	element	value
1	MX000017004	2010	1	30	tmax	278
2	MX000017004	2010	1	30	tmin	145
3	MX000017004	2010	2	2	tmax	273
4	MX000017004	2010	2	2	tmin	144
5	MX000017004	2010	2	3	tmax	241
6	MX000017004	2010	2	3	tmin	144
7	MX000017004	2010	2	11	tmax	297
8	MX000017004	2010	2	11	tmin	134
9	MX000017004	2010	2	23	tmax	299
10	MX000017004	2010	2	23	tmin	107
11	MX000017004	2010	3	5	tmax	321
12	MX000017004	2010	3	5	tmin	142
13	MX000017004	2010	3	10	tmax	345
14	MX000017004	2010	3	10	tmin	168
15	MX000017004	2010	3	16	tmax	311
16	MX000017004	2010	3	16	tmin	176
17	MX000017004	2010	4	27	tmax	363
18	MX000017004	2010	4	27	tmin	167
19	MX000017004	2010	5	27	tmax	332
20	MX000017004	2010	5	27	tmin	182

```
# dcast shifts variables from rows to columns
tidy <- dcast(raw1, ... ~ element)

# casting syntax:
#   row_var1 + row_var2 ~ col_var1 + col_var2
#   ... = all variables not otherwise mentioned
```

	id	year	month	day	tmax	tmin
1	MX000017004	2010	1	30	278	145
2	MX000017004	2010	2	2	273	144
3	MX000017004	2010	2	3	241	144
4	MX000017004	2010	2	11	297	134
5	MX000017004	2010	2	23	299	107
6	MX000017004	2010	3	5	321	142
7	MX000017004	2010	3	10	345	168
8	MX000017004	2010	3	16	311	176
9	MX000017004	2010	4	27	363	167
10	MX000017004	2010	5	27	332	182
11	MX000017004	2010	6	17	280	175
12	MX000017004	2010	6	29	301	180
13	MX000017004	2010	7	3	286	175
14	MX000017004	2010	7	14	299	165
15	MX000017004	2010	8	5	296	158
16	MX000017004	2010	8	8	290	173
17	MX000017004	2010	8	13	298	165
18	MX000017004	2010	8	23	264	150
19	MX000017004	2010	8	25	297	156
20	MX000017004	2010	8	29	280	153
21	MX000017004	2010	8	31	254	154
22	MX000017004	2010	10	5	270	140
23	MX000017004	2010	10	7	281	129
24	MX000017004	2010	10	14	295	130
25	MX000017004	2010	10	15	287	105

Tidy tools

Tidy tools

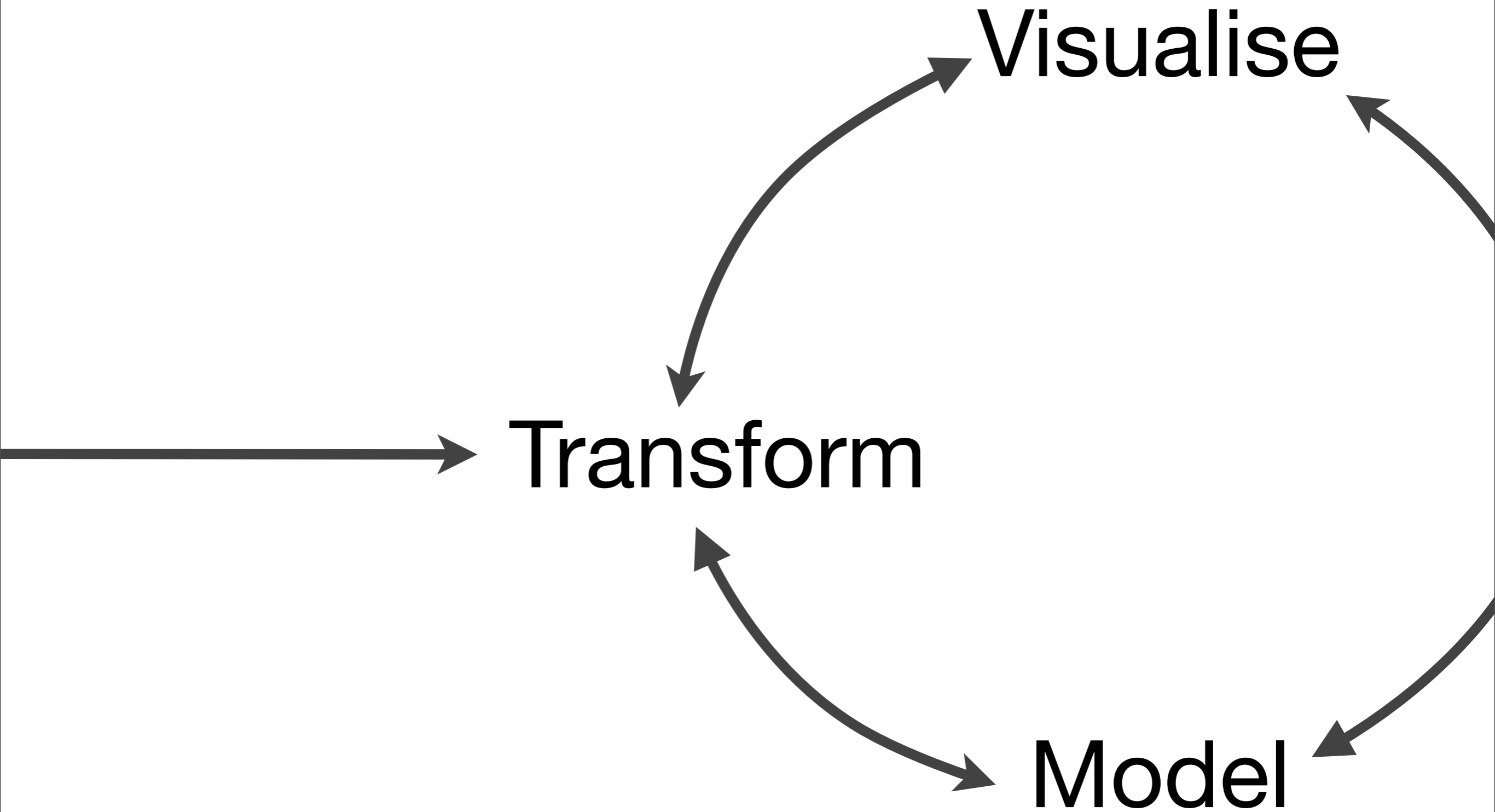
Now we have our data in a tidy format, what can we do with it?

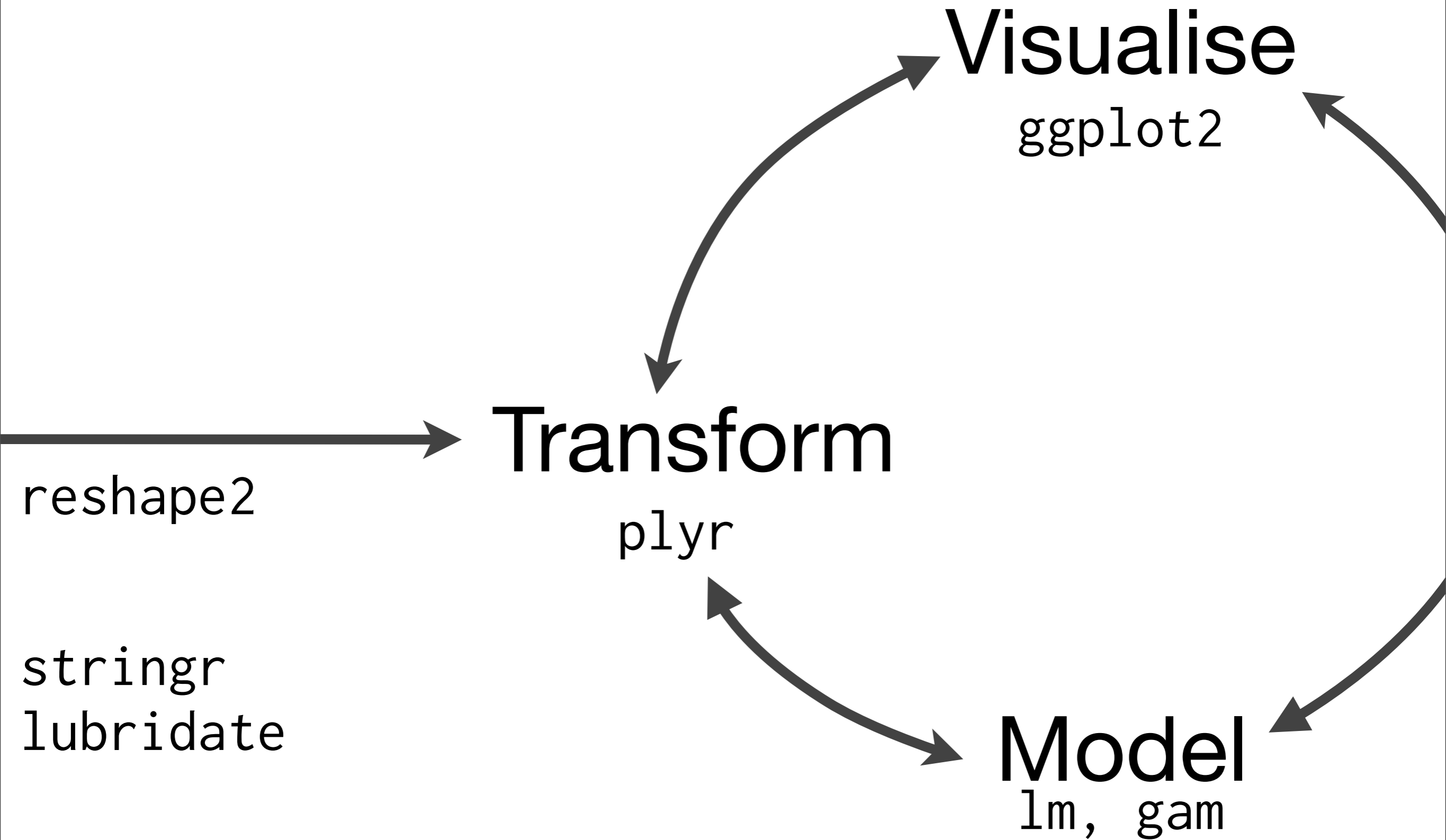
Tidy tools work input and output tidy data, and avoid data restructuring during an analysis.



<http://www.flickr.com/photos/wwworks/2473052504>

Monday, October 31, 11

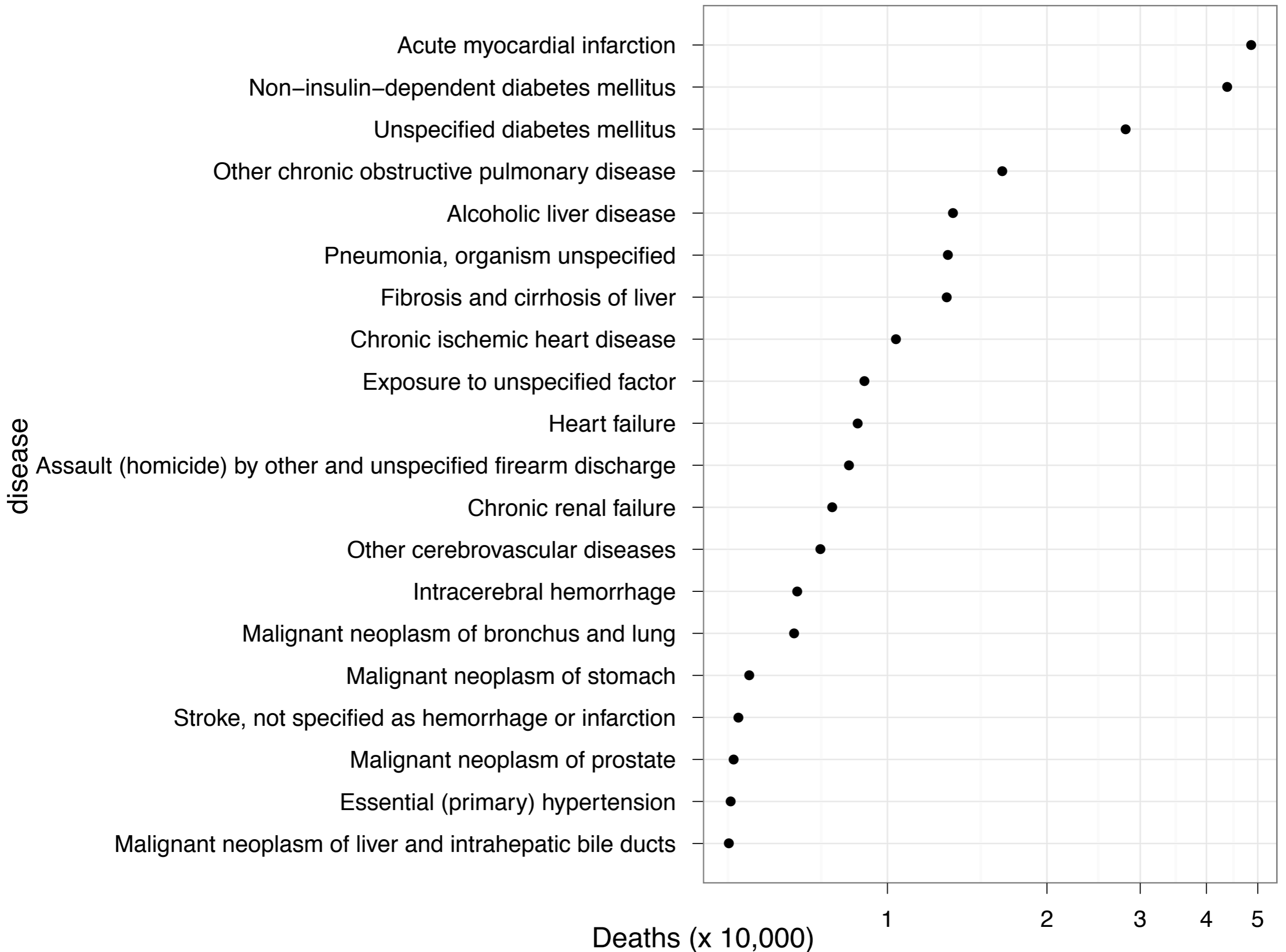






Function	Reason
<code>table()</code>	Returns an array
<code>by()</code>	Returns a list
<code>coef(summary())</code>	Returns a matrix with row names
<code>matplot()</code>	Inputs a matrix

Case study





What diseases don't fit this pattern?

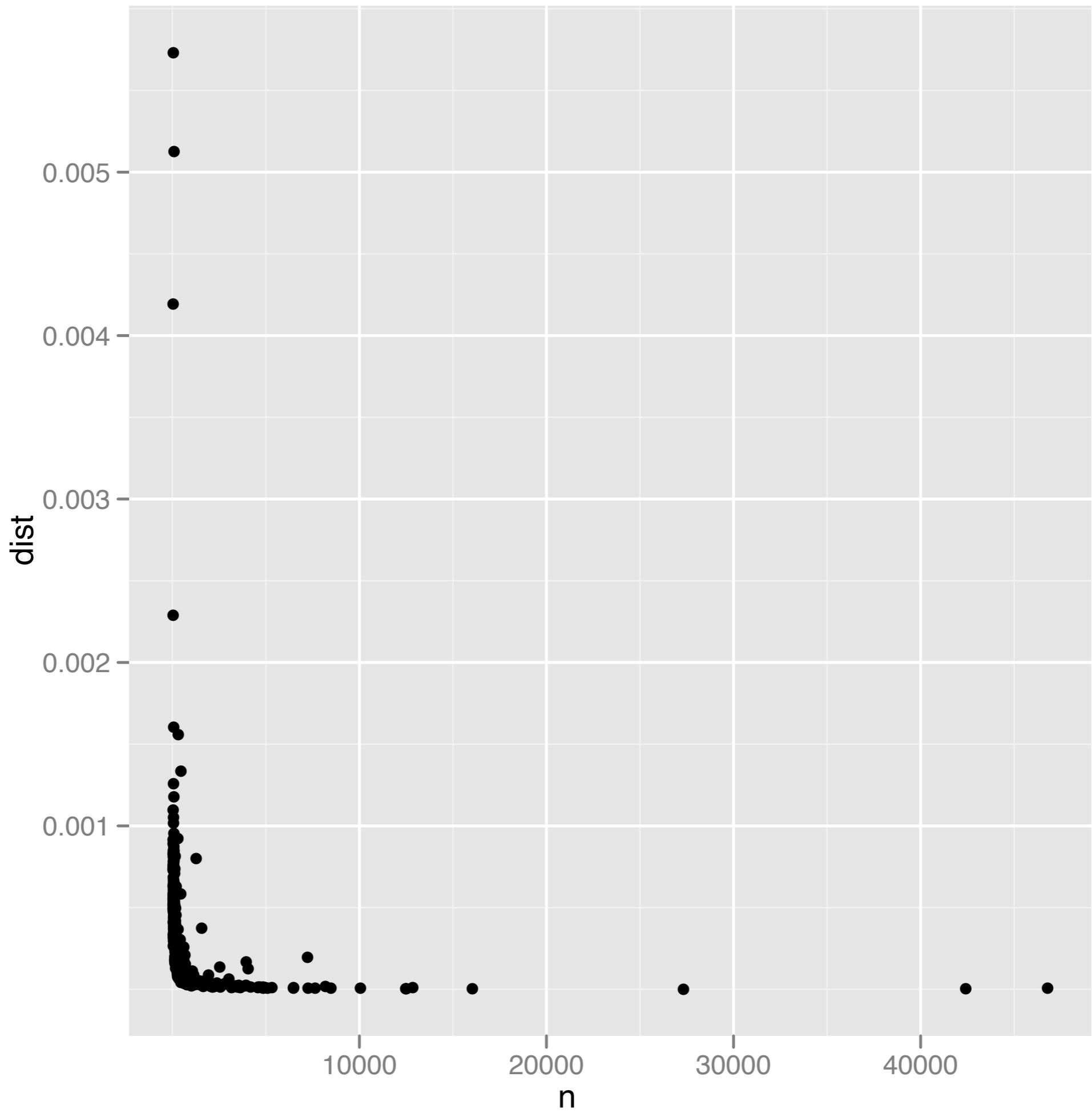
```
hod2 <- count(deaths, c("cod", "hod"))
hod2 <- subset(hod2, !is.na(hod))
hod2 <- join(hod2, codes)
hod2 <- ddply(hod2, "cod", transform,
  prop = freq / sum(freq))

# Compare to overall abundance
overall <- ddply(hod2, "hod", summarise,
  freq_all = sum(freq))
overall <- mutate(overall,
  prop_all = freq_all / sum(freq_all))

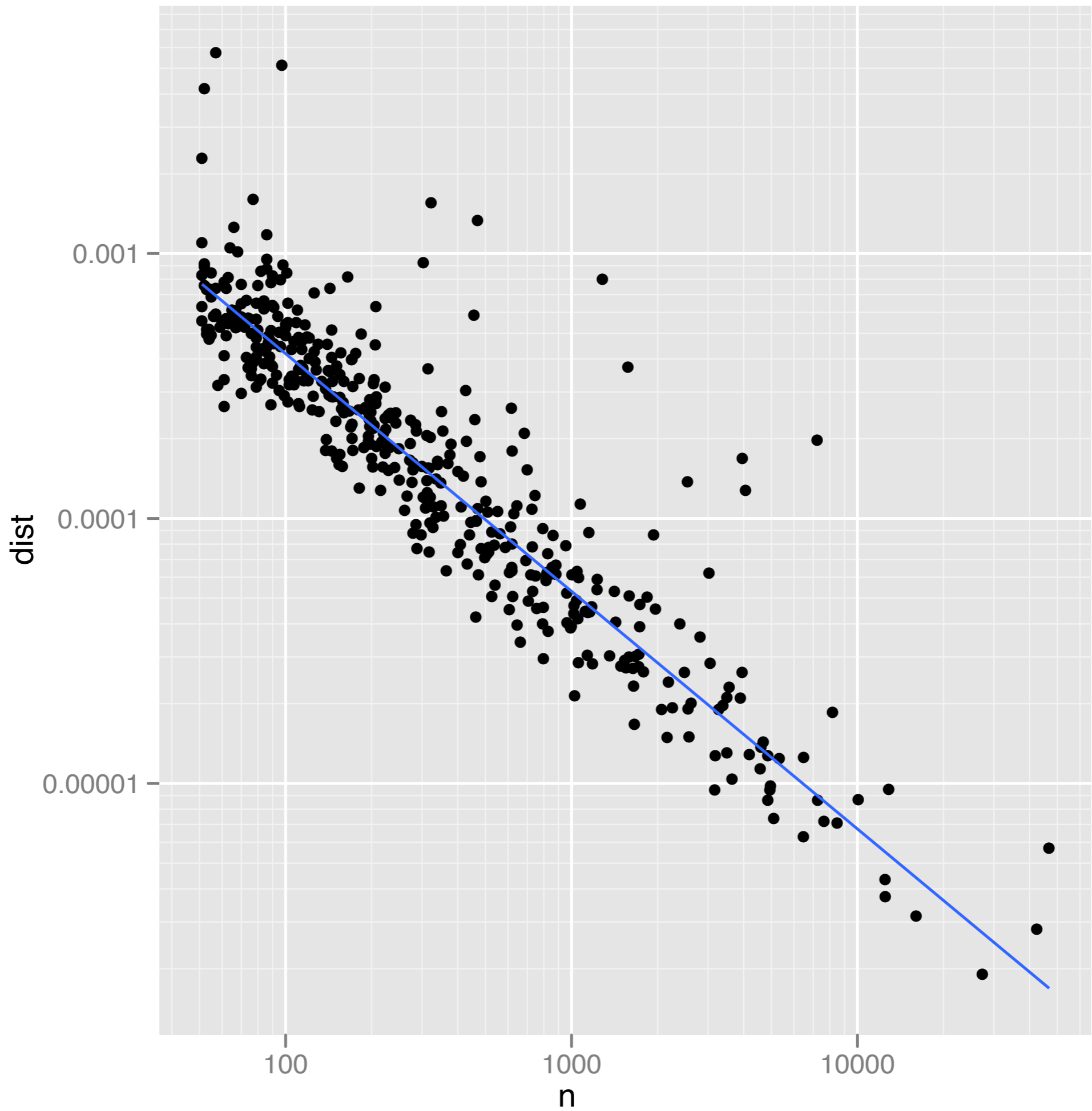
hod2 <- join(overall, hod2, by = "hod")
```


	cod	hod	disease	freq	prop	freq_all	prop_all
1	A01	1	Typhoid and paratyphoid\nfevers	3	0.0577	20430	0.0398
2	A01	2	Typhoid and paratyphoid\nfevers	1	0.0192	18962	0.0369
3	A01	3	Typhoid and paratyphoid\nfevers	4	0.0769	19729	0.0384
4	A01	5	Typhoid and paratyphoid\nfevers	5	0.0962	22126	0.0431
5	A01	6	Typhoid and paratyphoid\nfevers	1	0.0192	23787	0.0463
6	A01	8	Typhoid and paratyphoid\nfevers	1	0.0192	21915	0.0427
7	A01	10	Typhoid and paratyphoid\nfevers	2	0.0385	24321	0.0474
8	A01	11	Typhoid and paratyphoid\nfevers	2	0.0385	23843	0.0465
9	A01	12	Typhoid and paratyphoid\nfevers	1	0.0192	23392	0.0456
10	A01	13	Typhoid and paratyphoid\nfevers	6	0.1154	23284	0.0454
11	A01	14	Typhoid and paratyphoid\nfevers	4	0.0769	23053	0.0449
12	A01	15	Typhoid and paratyphoid\nfevers	5	0.0962	23278	0.0454
13	A01	17	Typhoid and paratyphoid\nfevers	3	0.0577	23625	0.0460
14	A01	18	Typhoid and paratyphoid\nfevers	2	0.0385	24380	0.0475
15	A01	19	Typhoid and paratyphoid\nfevers	3	0.0577	22919	0.0447
16	A01	20	Typhoid and paratyphoid\nfevers	3	0.0577	22926	0.0447
17	A01	21	Typhoid and paratyphoid\nfevers	2	0.0385	20995	0.0409
18	A01	22	Typhoid and paratyphoid\nfevers	3	0.0577	20510	0.0400
19	A01	23	Typhoid and paratyphoid\nfevers	1	0.0192	21446	0.0418

```
devi <- ddply(hod2, "cod", summarise, n = sum(freq),  
  dist = mean((prop - prop_all)^2))  
devi <- subset(devi, n > 50)  
  
qplot(n, dist, data = devi)
```

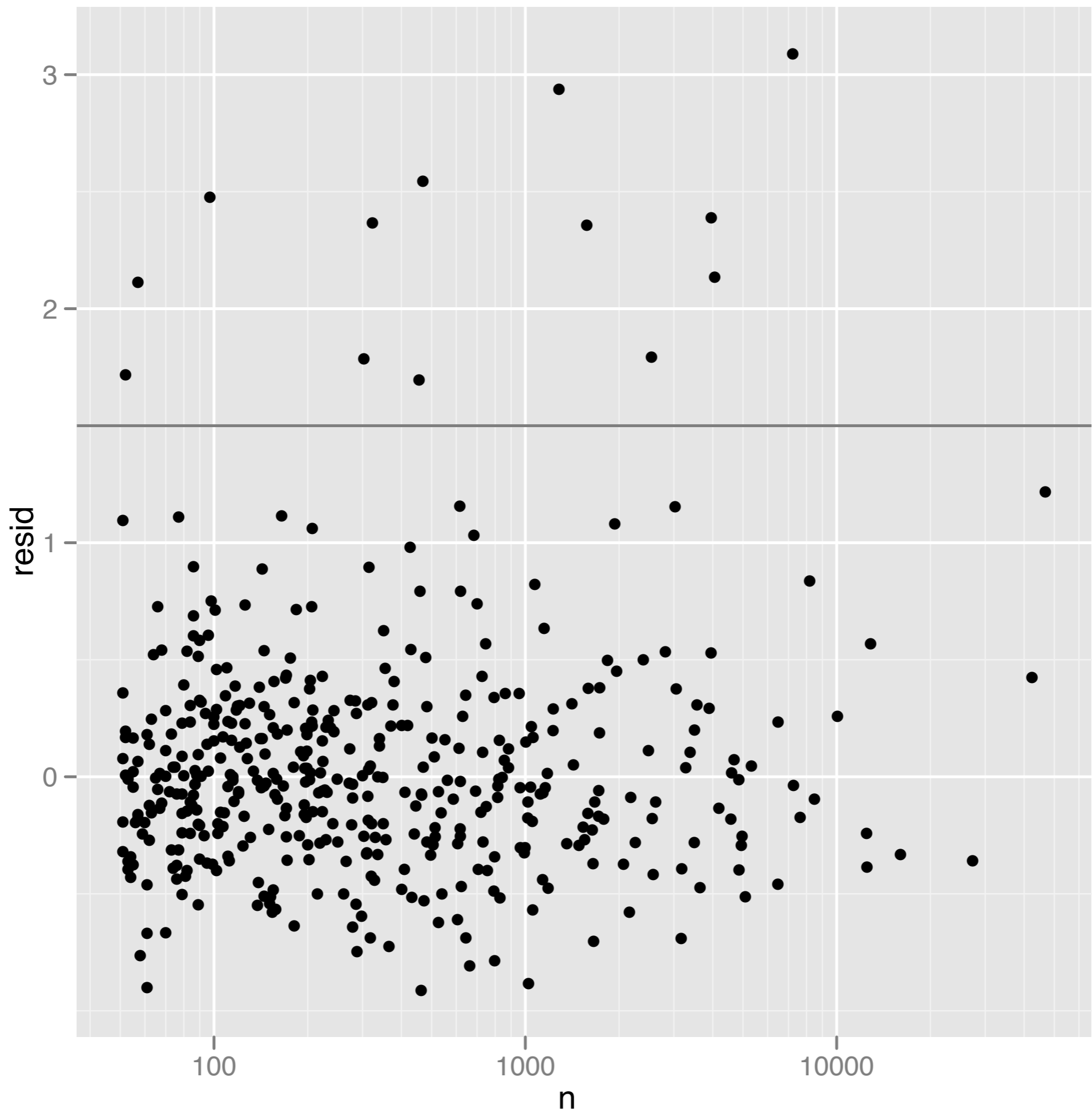


```
qplot(n, dist, data = devi) +  
  geom_smooth(method = "rlm", se = F) +  
  xlog10 +  
  ylog10  
  
xlog10 <- scale_x_log10(  
  breaks = c(100, 1000, 10000),  
  labels = c(100, 1000, 10000),  
  minor_breaks = outer(1:9, 10^(1:5), "*"))  
ylog10 <- scale_y_log10(  
  breaks = 10 ^ -c(3, 4, 5),  
  labels = c("0.001", "0.0001", "0.00001"),  
  minor_breaks = outer(1:9, 10^-(3:6), "*"))
```



```
devi$resid <- resid(rlm(log(dist) ~ log(n),
  data = devi))

ggplot(devi, aes(n, resid)) +
  geom_hline(yintercept = 1.5, colour = "grey50") +
  geom_point() +
  xlog10
```

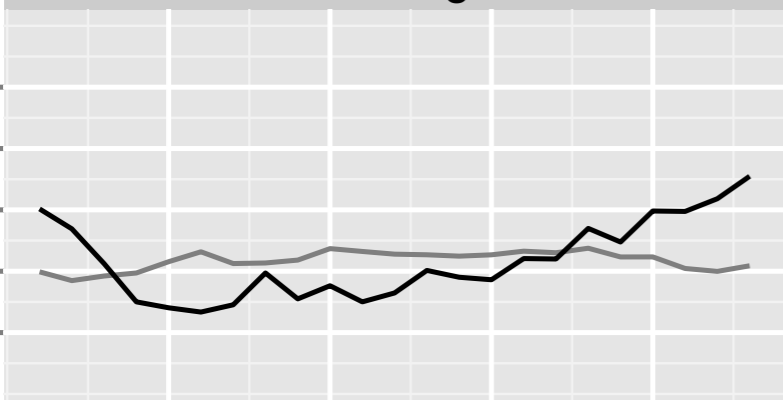


```
unusual <- subset(devi, resid > 1.5)
hod_unusual_big <- match_df(hod2, subset(unusual, n > 350))
hod_unusual_sml <- match_df(hod2, subset(unusual, n <= 350))

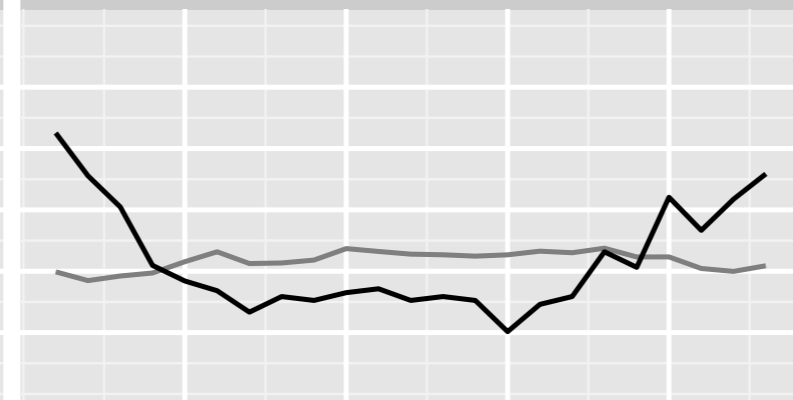
# Visualise unusual causes of death
ggplot(hod_unusual_big, aes(hod, prop)) +
  geom_line(aes(y = prop_all), data = overall, colour = "grey50") +
  geom_line() +
  facet_wrap(~ disease, ncol = 3)
```


Assault (homicide) by other and unspecified firearm discharge

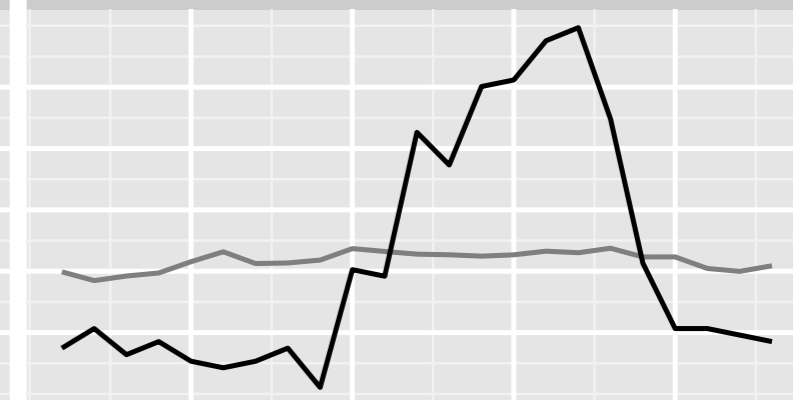
0.10
0.08
0.06
0.04
0.02



Assault (homicide) by sharp object



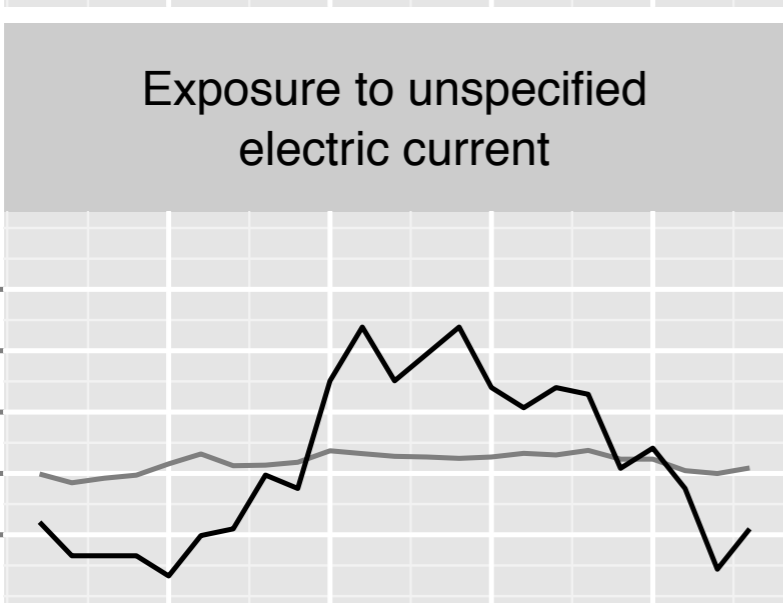
Drowning and submersion while in natural water



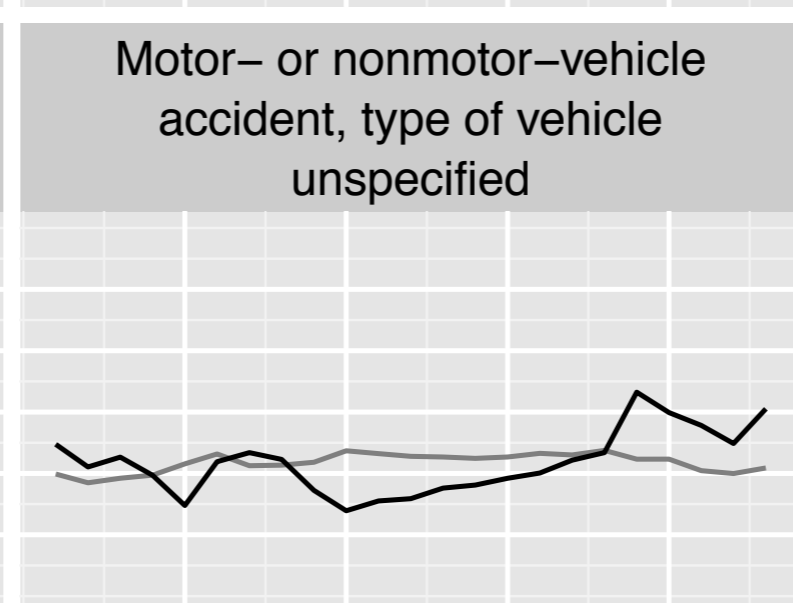
Exposure to unspecified electric current

prop

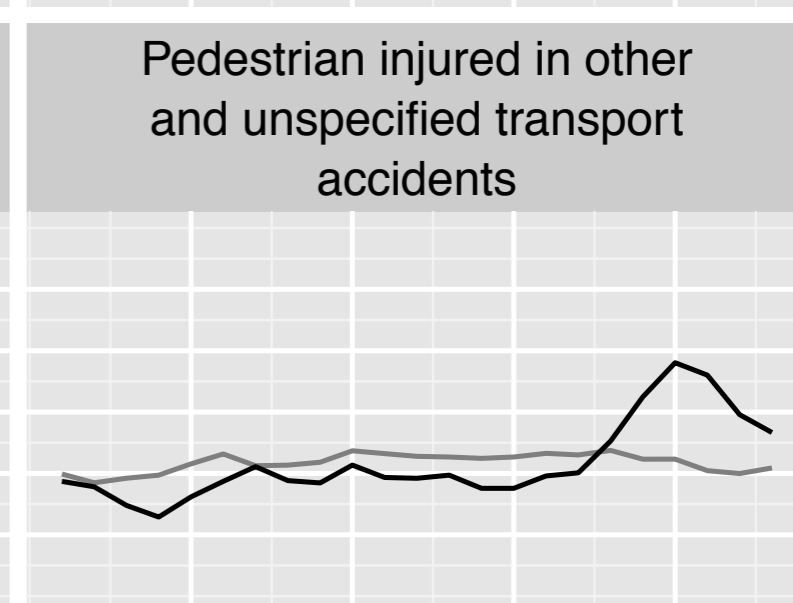
0.10
0.08
0.06
0.04
0.02



Motor- or nonmotor-vehicle accident, type of vehicle unspecified

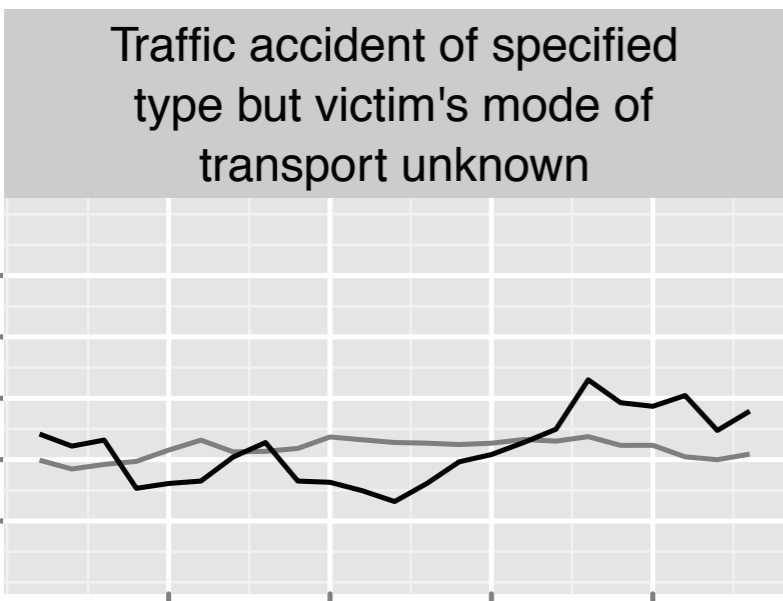


Pedestrian injured in other and unspecified transport accidents

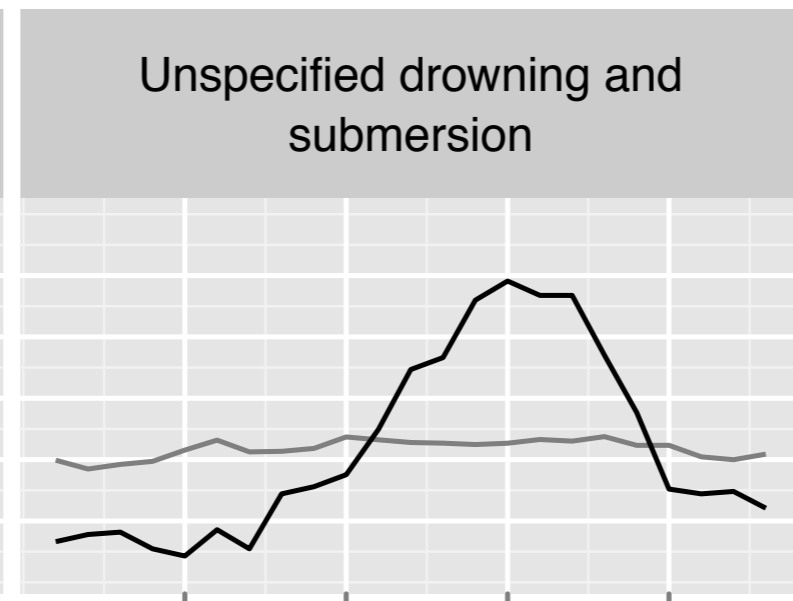


Traffic accident of specified type but victim's mode of transport unknown

0.10
0.08
0.06
0.04
0.02



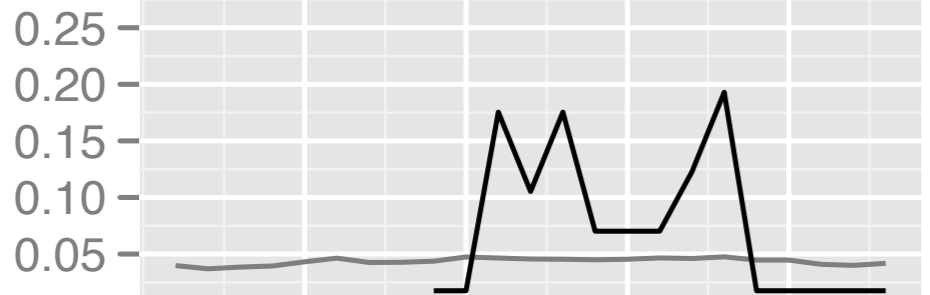
Unspecified drowning and submersion



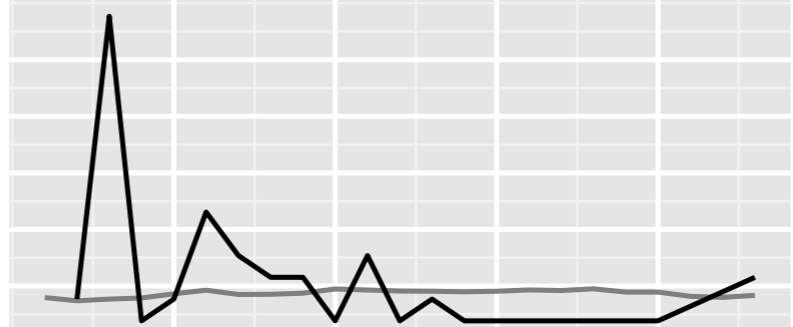
hod

5 10 15 20

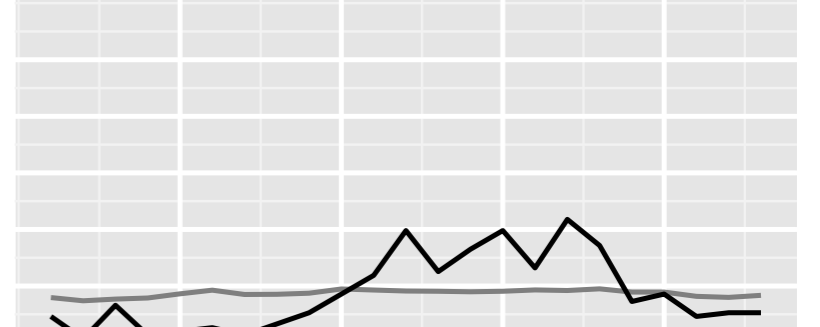
Accident to powered aircraft causing injury to occupant



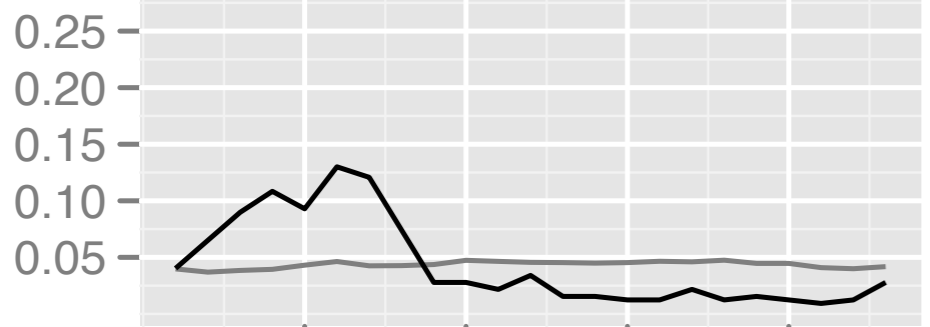
Bus occupant injured in other and unspecified transport accidents



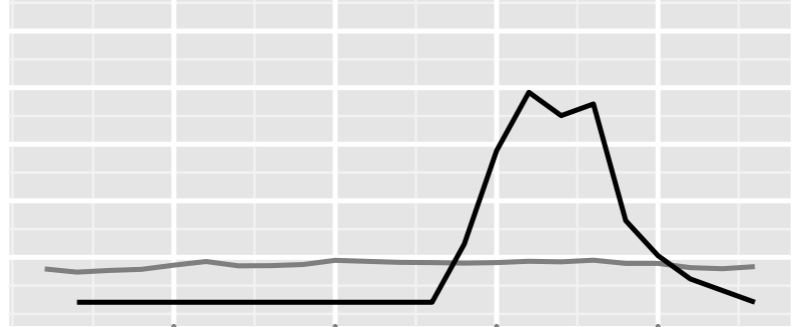
Other specified drowning and submersion



Sudden infant death syndrome



Victim of lightning



hod

prop

Conclusions

Summary

The framework of tidy data makes it easier to get data in a useful form for analysis and provides a useful framework for critiquing existing functions.

Surprisingly few tools needed to tidy messy data.

Future work

Data structure also affects how we think about problem statistically:

Multivariate models use matrices

Paired t-test vs. mixed effect model

```
library(lme4); set.seed(1001)

x <- rnorm(10, 20, 1)
df <- data.frame(
  id = 1:10,
  x = x,
  y = x + rnorm(10, 2, 1))

# Paired t-test directly
t1 <- with(df, t.test(x, y, paired = TRUE))

# With mixed model (courtesy of Ben Bolker)
dfm <- melt(df, "id")
m1 <- lmer(value ~ variable + (1 | id), data = dfm, REML = T)

all.equal(
  abs(t1$statistic),
  coef(summary(m1))["variabley", "t value"])
```

<http://vita.had.co.nz/papers.html>
<http://vita.had.co.nz/presentations.html>