MONASH University

Data Wrangling

Di Cook
Professor of Business Analytics

Motivation

- What can you do if you have some basic data wrangling skills?
- A few examples……

US Airline traffic

- ~15,000 flights a day
- April 1986 - present
- RITA - Research and Innovative Technology Administration (flight information, arrival delay, airline, plane id, …)
- On time performance database - http://www.transtats.bts.gov/

You can download this yourself!

Average arrival delay, minutes

One year

20

-5

midnight

midnight

Average Arrival Delay

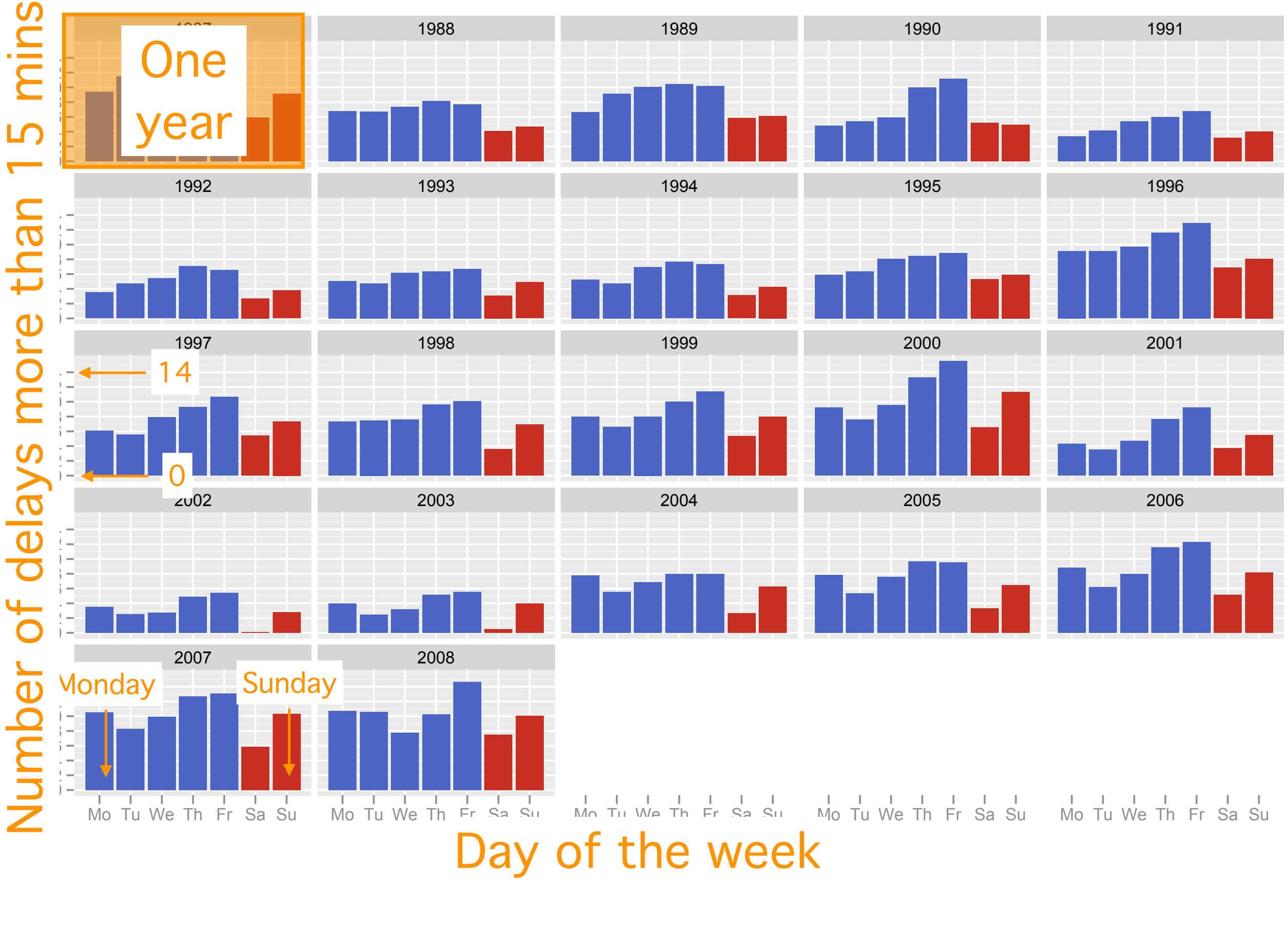
1987 1988 1989 1990 1991
1992 1993 1994 1995 1996
1997 1998 1999 2000 2001
2002 2003 2004 2005 2006
2007 2008

20 15 10 5 0 -5

500 1000 1500 2000

Year
1990
1995
2000
2005

Scheduled departure time, local time, by hour


Number of delays more than 15 mins
One year
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
14
0
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
Monday
Sunday
2008
Mo Tu We Th Fr Sa Su
Day of the week

Arrival delay (min)

EWR
SFO
LGA
ORD
JFK

Airports (worst 50 by delays)

DFW

Destination

EWR SFO LGA ORD JFK PHL FLL BOS SEA CMH ATL MIA TPA PDX IND LAX LAS SAN SMF DEN MCO MCI MSY HNL BWI ONT SAT CLE PIT RDU ABQ PHX IAD AUS IAH SLC OAK MDW STL HOU DFW DCA CVG DAL CLT SNA DTW SJC BNA MSP MEM

0 5 10 15 20

Arrival Delay (min)

Average Daily Flights

ORD

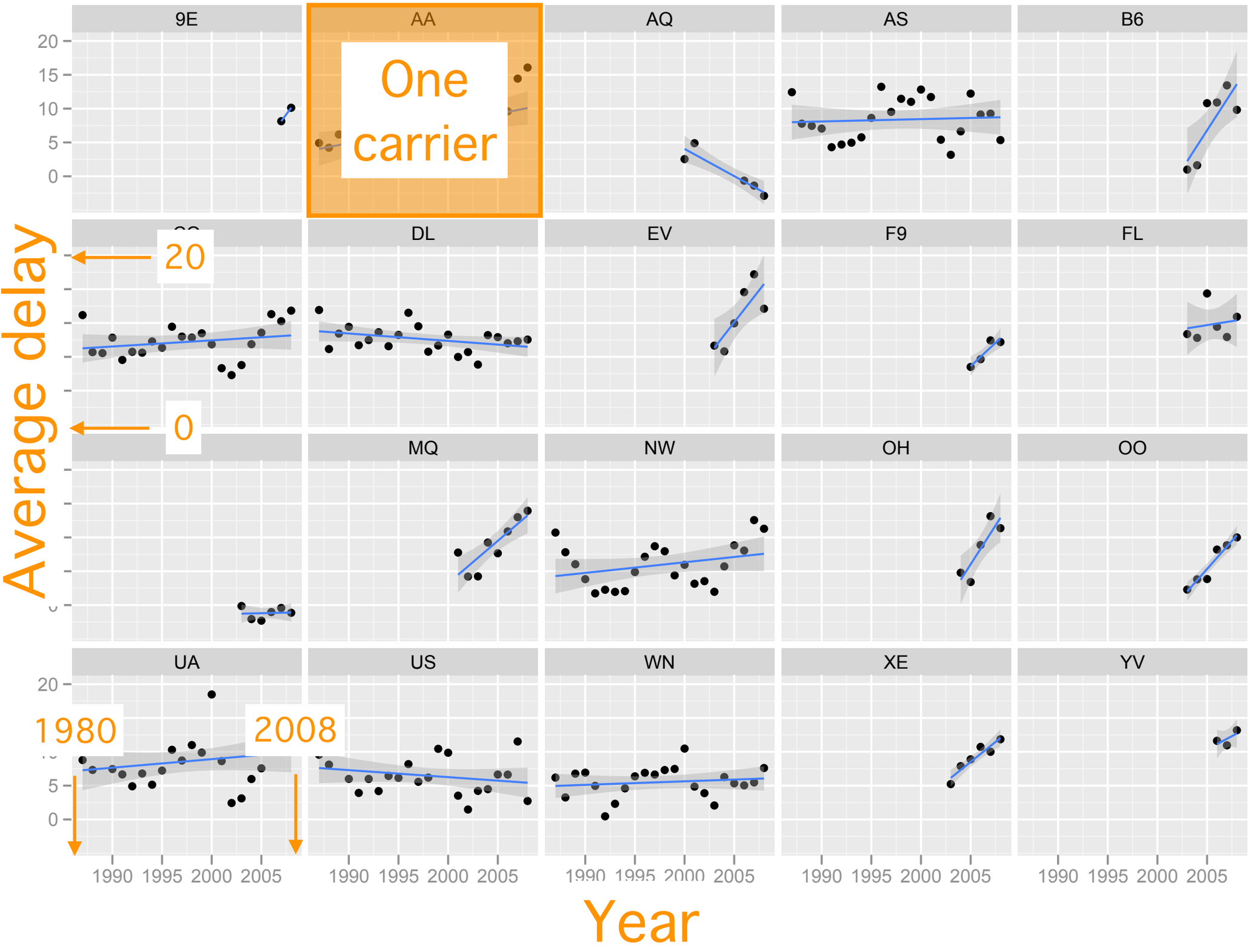
DFW

0 200 400 600

count

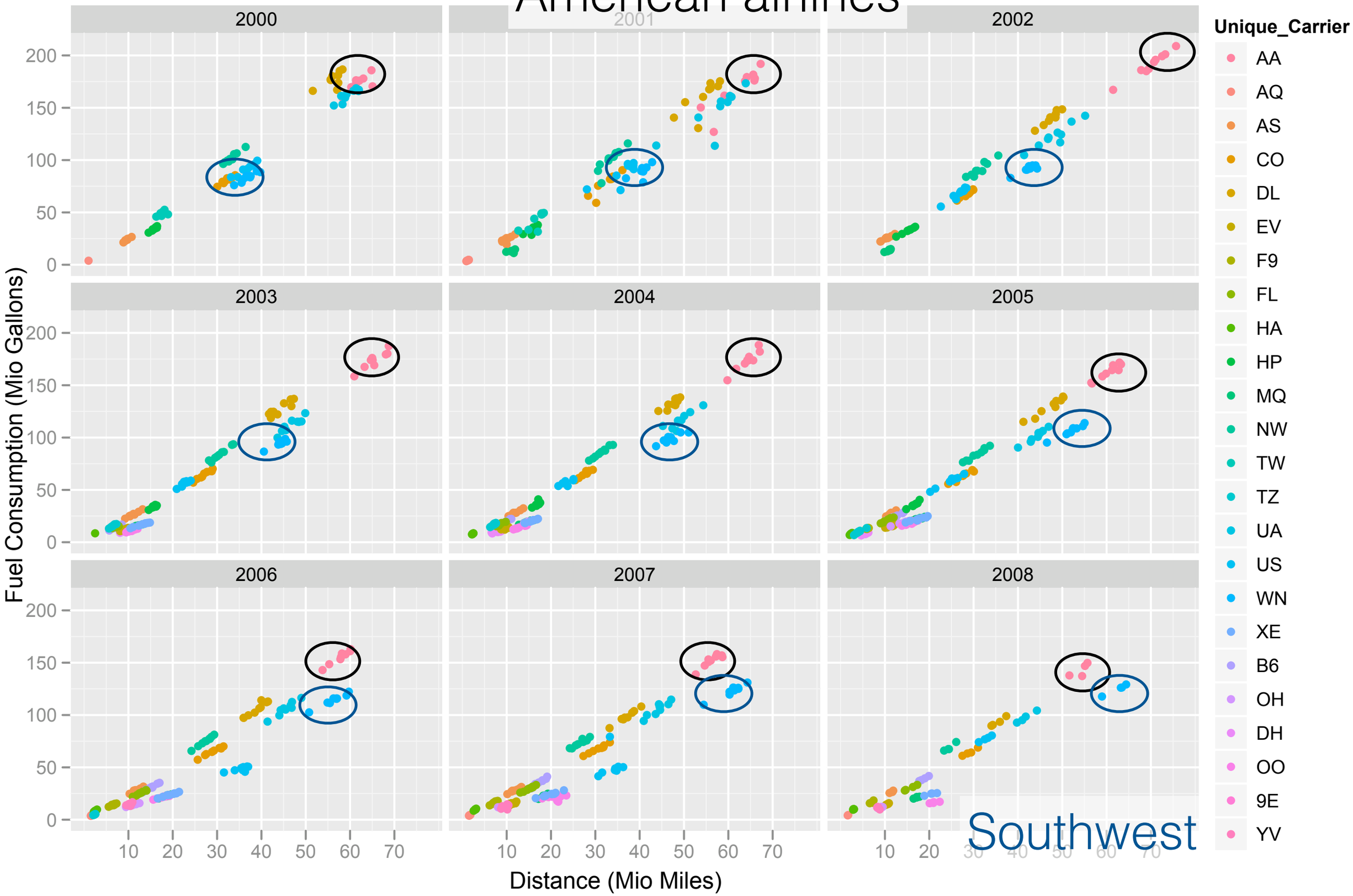
How we can use this

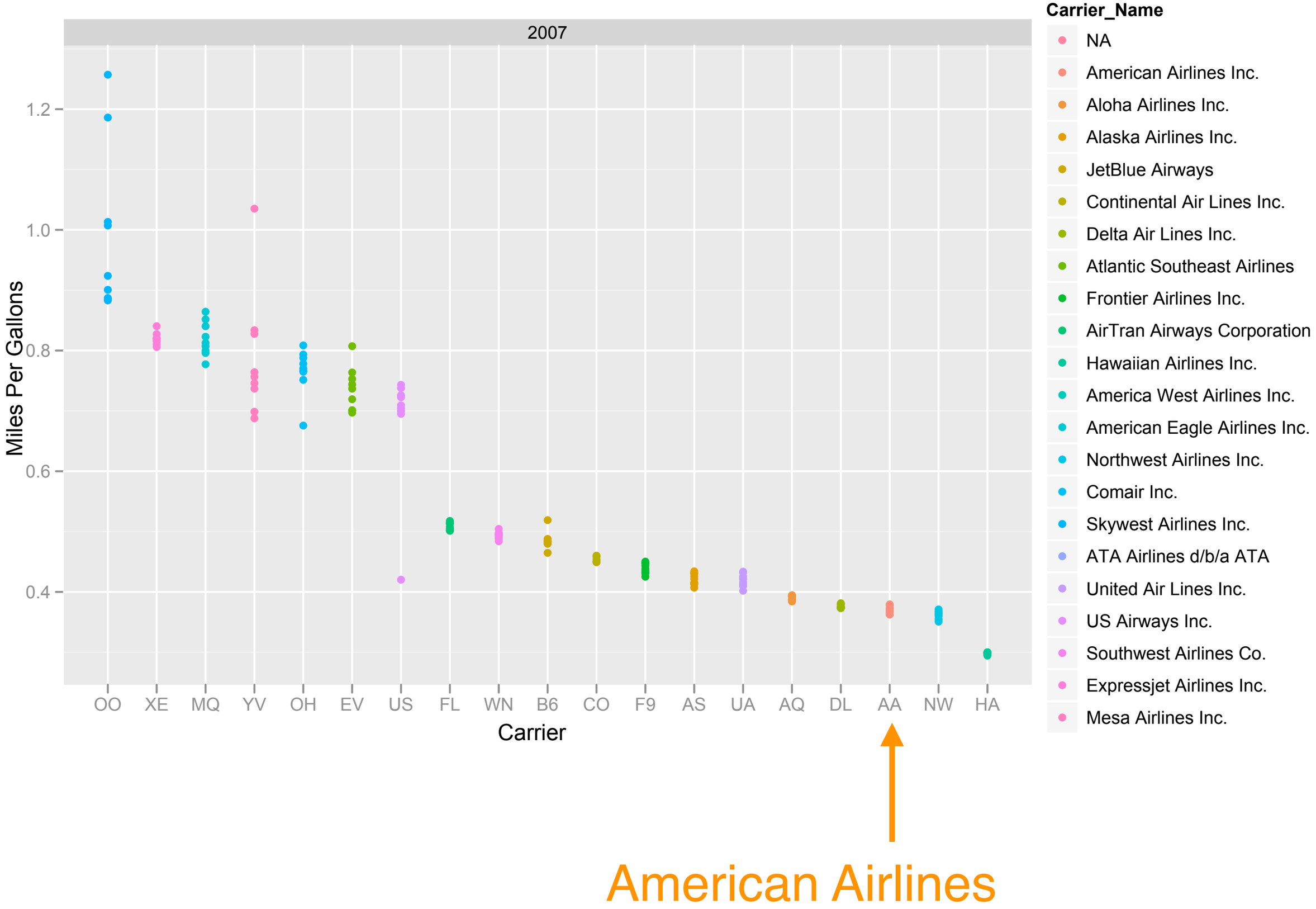
- Fly early in the day, early in the week or weekends (Saturday)
- Avoid ORD, JFK, LGA, EWR
- What about carriers? General events?



Carrier	Name
9E	Pinnacle
AA	American
AQ	Aloha
AS	Alaska
B6	JetBlue
CO	Continental
DL	Delta
EV	Atlantic Southeast
F9	Frontier
FL	AirTran
HA	Hawaiian
HP	America West
MQ	American Eagle
NW	Northwest
OH	Comair
OO	Skywest
UA	United
US	US Airways
WN	Southwest
XE	Expressjet
YV	Mesa

American airlines





American Airlines

American Airlines

- American airlines filed for bankruptcy Nov 29, 2013
- Mining publicly available data could have sounded the alarms several years in advance

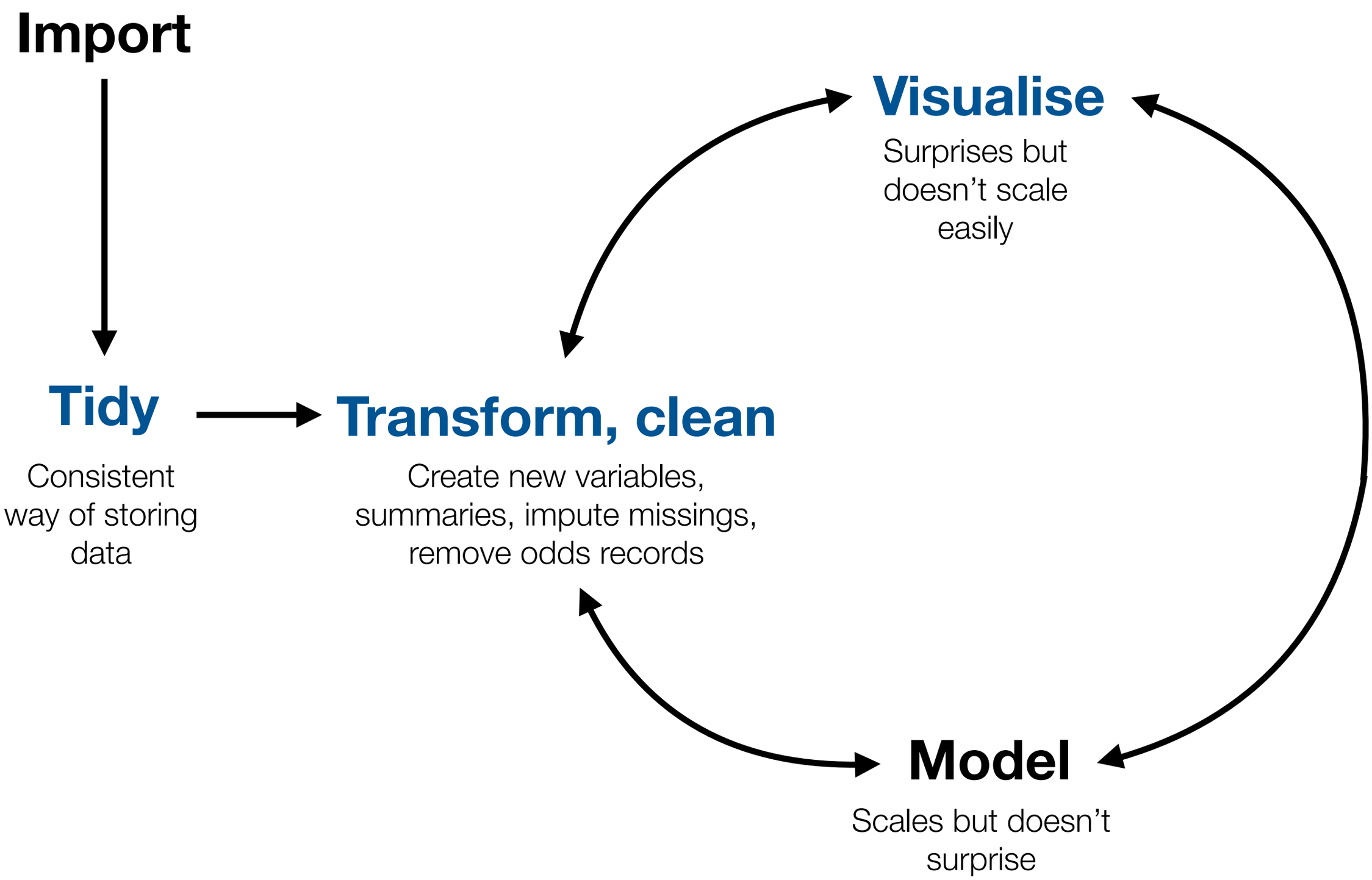


Diagram from Hadley Wickham

Raw data

First 5 records

	FL_DATE	CARRIER	TAIL_NUM	FL_NUM	ORIGIN	ORIGIN_ST	DEST	DEST_ST
1	2015-08-07	UA	N14219	1650	LAS	NV	ORD	IL
2	2015-08-07	UA	N69830	1650	ORD	IL	SEA	WA
3	2015-08-07	UA	N76529	1652	BWI	MD	ORD	IL
4	2015-08-07	UA	N76529	1652	DEN	CO	BWI	MD
5	2015-08-07	UA	N37465	1652	EWR	NJ	DEN	CO

CRS_DEP_TIME	DEP_TIME	DEP_DELAY	CRS_ARR_TIME	ARR_TIME	ARR_DELAY	CRS_EL_TIME	ACT_EL_TIME	AIR_TIME	DIST
1400	1355	-5.00	1940	0239		220.00			1514.00
2245	2240	-5.00	0110	0055	-15.00	265.00	255.00	235.00	1721.00
1723	1737	14.00	1830	1835	5.00	127.00	118.00	91.00	622.00
1110	1109	-1.00	1638	1620	-18.00	208.00	191.00	172.00	1491.00
806	804	-2.00	1022	1004	-18.00	256.00	240.00	214.00	1605.00

http://bit.ly/wrangling1

Your Turn

What are the (1) rows, (2) columns of the data?

Answers:

(1) Rows contain the information about one flight that has occurred in the USA

(2) Columns have information about that flight such as the carrier, origin, destination, date, scheduled departure time, actual departure time, flight time, distance

Types of variables

- Quantitative: Numerical values that can be ordered
  - Continuous: can take any real-value
  - Discrete: separated set of values, eg integers
- Categorical: Information that can be divided into groups
  - Ordinal: categories can be ordered
  - Nominal: no natural order to categories
- Temporal: Time variable, date, year, month, week, …
- Spatial: Geographic location, latitude, longitude

Types of variables

- Variable type determines what cleaning can be done, want analysis is appropriate, how to plot it…
- Understanding what type of information is available is important

http://bit.ly/wrangling2

Your Turn

What are the types of variables in the airline data?

Answers:

(1) Quantitative, continuous; Quantitative, discrete; Categorical, nominal; Temporal

(2) None of them!

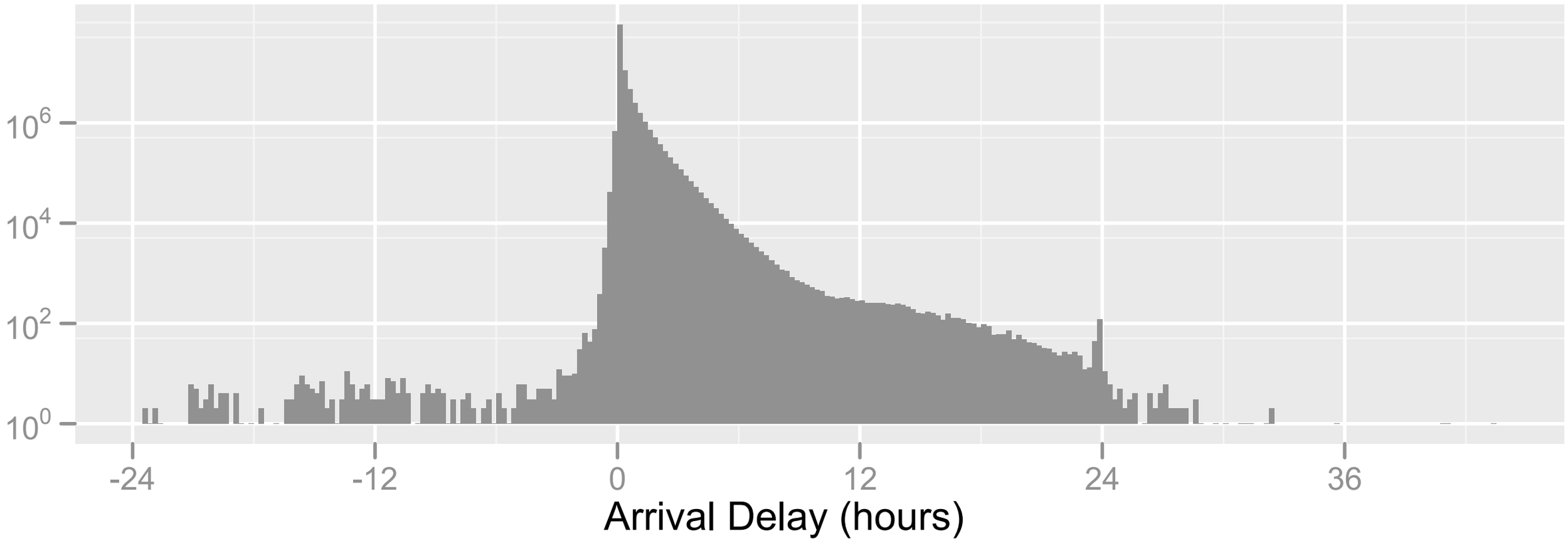
RIDDLES

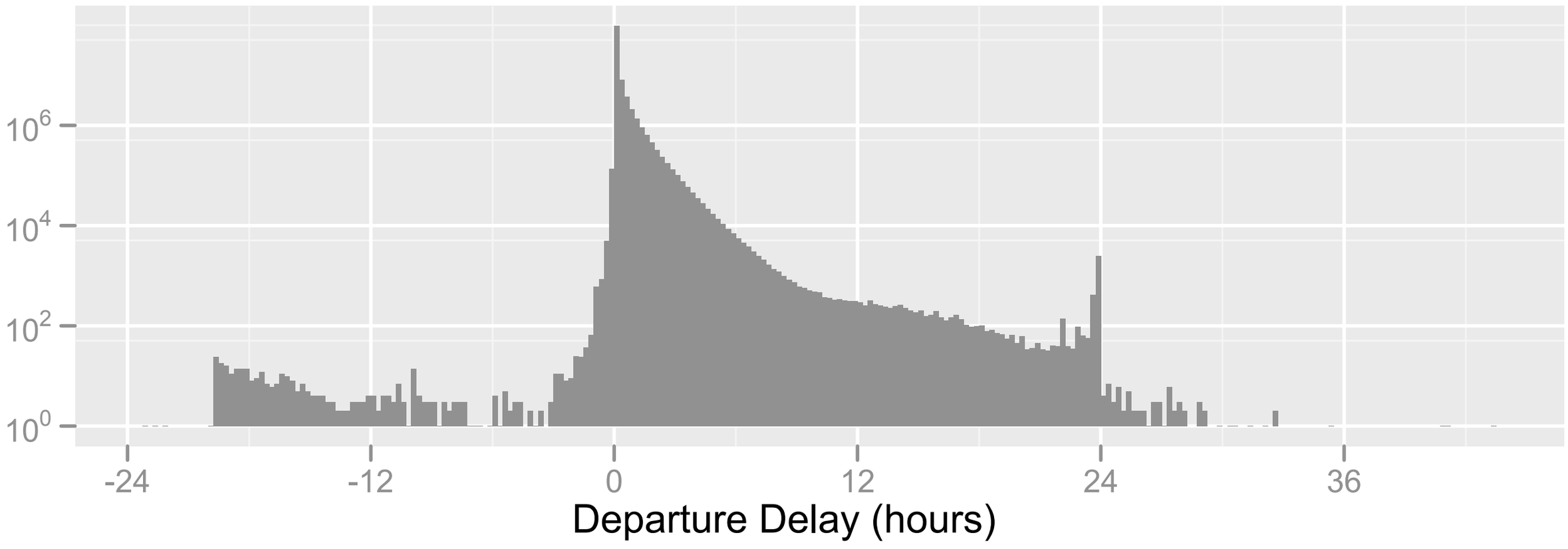
When is age not a quantitative variable?

When is educational level not a categorical variable?

Exploring delays

- Question: Do delays change in the course of the day? Over the years? By carrier? By airport?
- What needs to be done to the data?
- Delay is a quantitative variable given in minutes
- Data is on database, if you pull everything back you will have 18Gb of information to proces
- Make the database work - minimally - aggregate delays into bins - produce a histogram

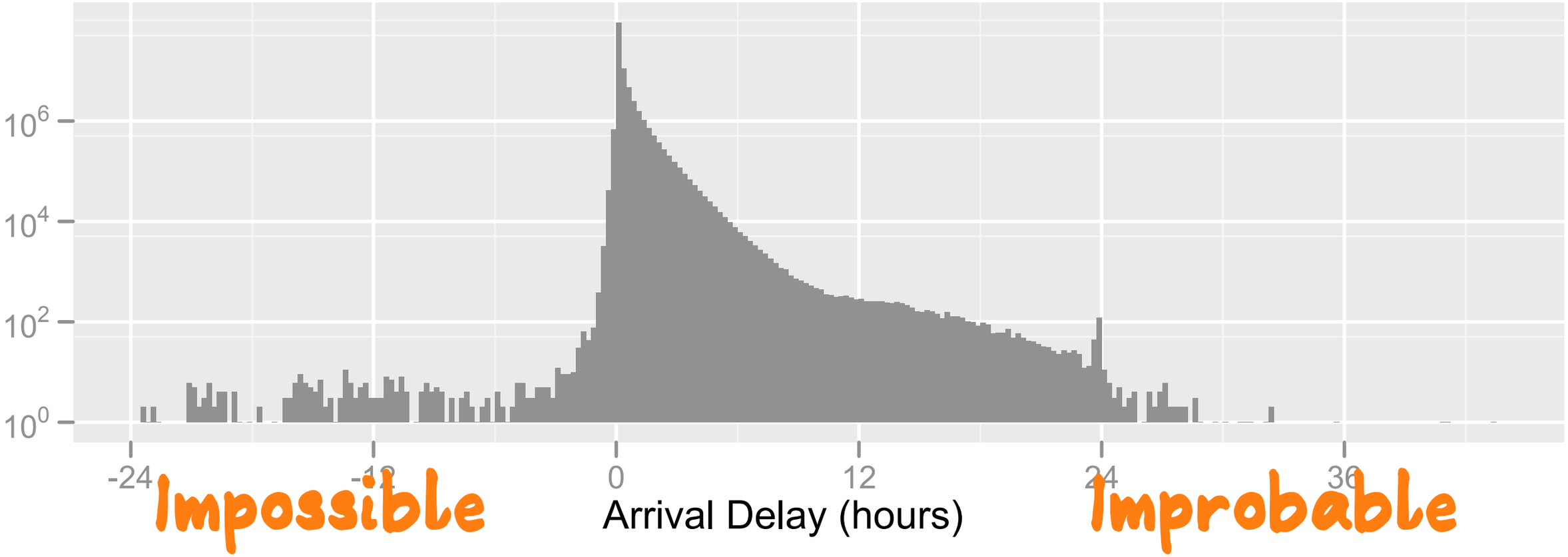

10^6
10^4
10^2
10^0
-24
-12
0
12
24
36
Arrival Delay (hours)

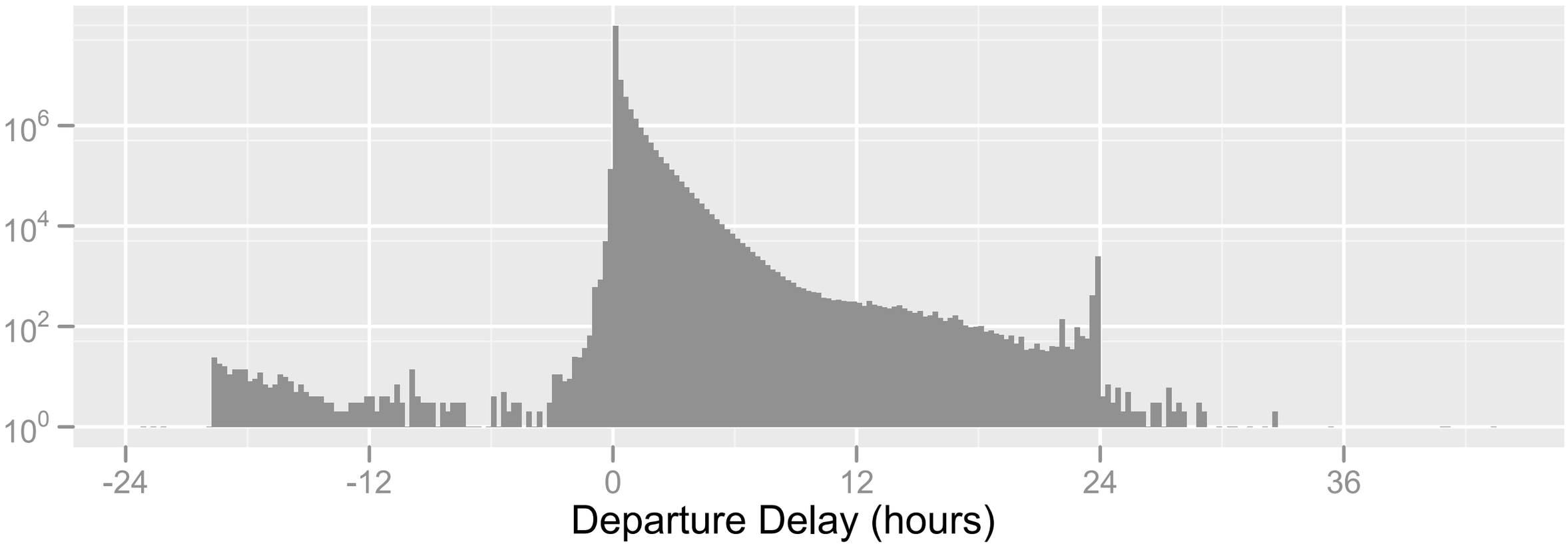

10^6
10^4
10^2
10^0
-24
-12
0
12
24
36
Departure Delay (hours)

Delays

- Arrival delays range from -24 hours to 30 hours
- Departure delays range from -18 hours to 30 hours
- Peak is at 0, no delay
- Secondary peaks at 24 hours

Really?


10^6
10^4
10^2
10^0
-24
-12
0
12
24
36
Impossible
Arrival Delay (hours)
Improbable


10^6
10^4
10^2
10^0
-24
-12
0
12
24
36
Departure Delay (hours)

Reasons?

- Delay calculated incorrectly for some flights
- Mistakes in arrival or departure times, leading to mistakes in the delay values
- Failure to convert local time to global time
- ….
- What do you do?

Detection and correction

Your Turn

Brainstorm with your neighbor(s) and come up with **three** questions that you'd like to ask about airline traffic

Your Turn

Pick one of the questions,
brainstorm with your neighbor(s)
what steps you would need to take
in order to answer the question

Here's one

- Track the movements of plane XXX (N478HA)
- We need more information: lat/long of airports (also found at BTS web site)

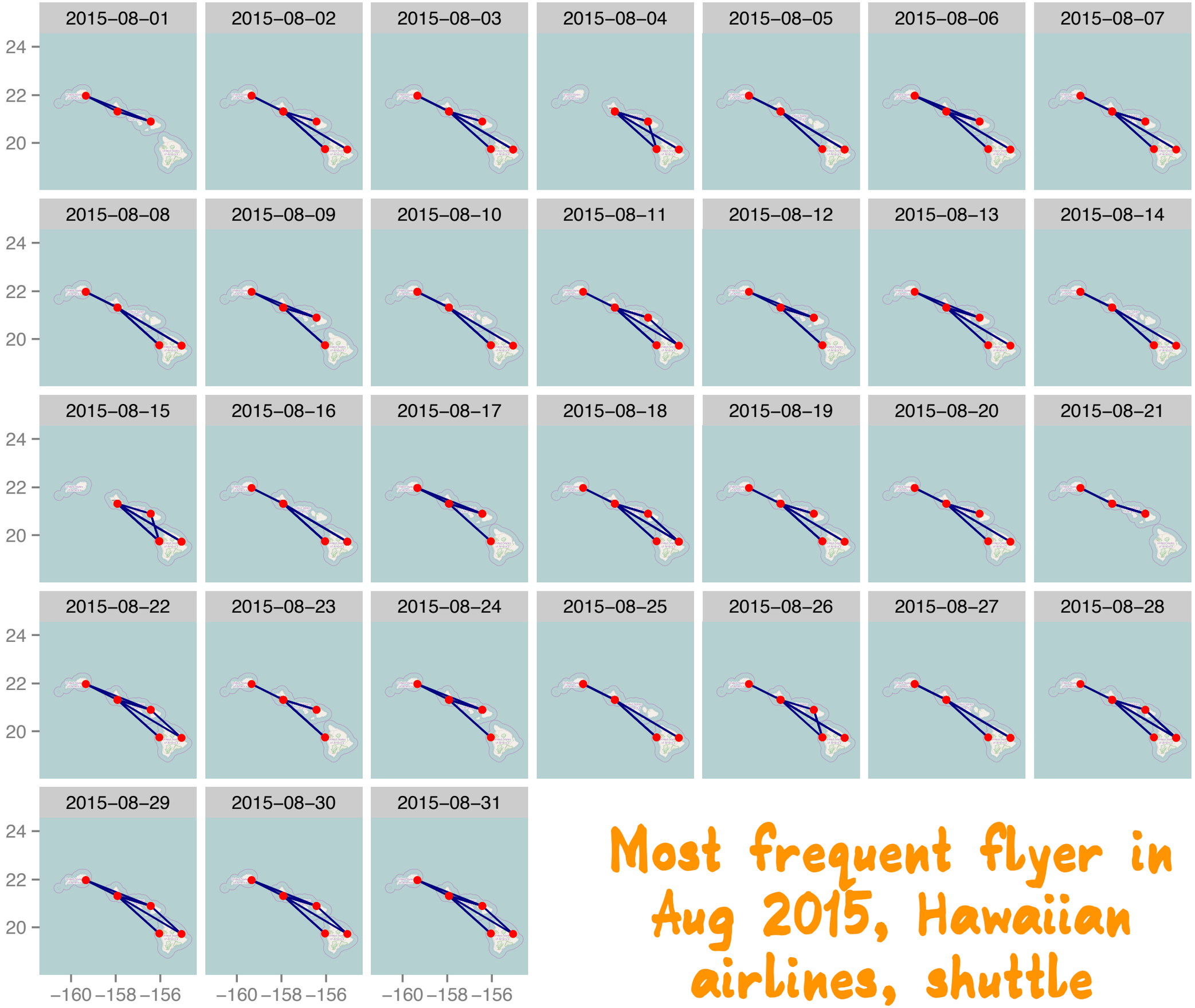
	FL_DATE	CARRIER	FL_NUM	ORIGIN	DEST	DEP_TIME	ARR_TIME	DISTANCE
1	2015-08-01	HA	156	HNL	OGG	704	746	100
2	2015-08-01	HA	155	OGG	HNL	814	850	100
3	2015-08-01	HA	174	HNL	OGG	924	1010	100
4	2015-08-01	HA	211	OGG	LIH	1037	1122	201
5	2015-08-01	HA	144	LIH	HNL	1155	1229	102
6	2015-08-02	HA	123	HNL	LIH	721	759	102

Flights

	AIRPORT	LATITUDE	LONGITUDE
1	AHS	15.47278	-84.353056
2	AHT	51.37861	179.258611
3	AHU	35.17722	-3.839444
4	AIA	42.05333	-102.803611
5	AIB	56.19000	-132.445833
6	AID	40.10861	-85.613056
7	AIK	33.64944	-81.685000
8	AIN	70.63806	-159.994722
9	AIT	-18.83750	-159.761111
10	OGG	20.89861	-156.4306

Airports

Linking information from different sources



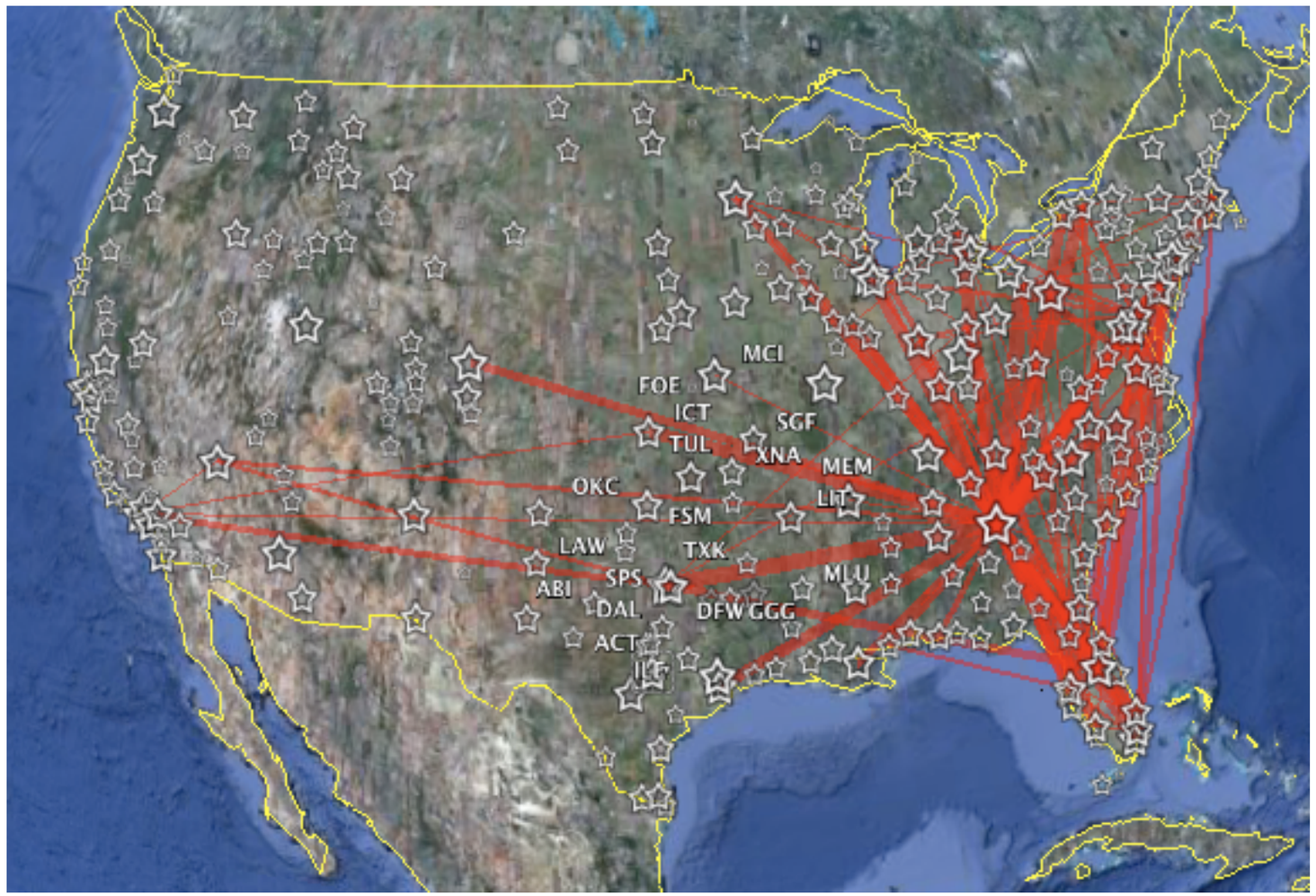
Most frequent flyer in Aug 2015, Hawaiian airlines, shuttle

Expanding

- Look at gaps in the records - flight with NO passengers - GHOST flights

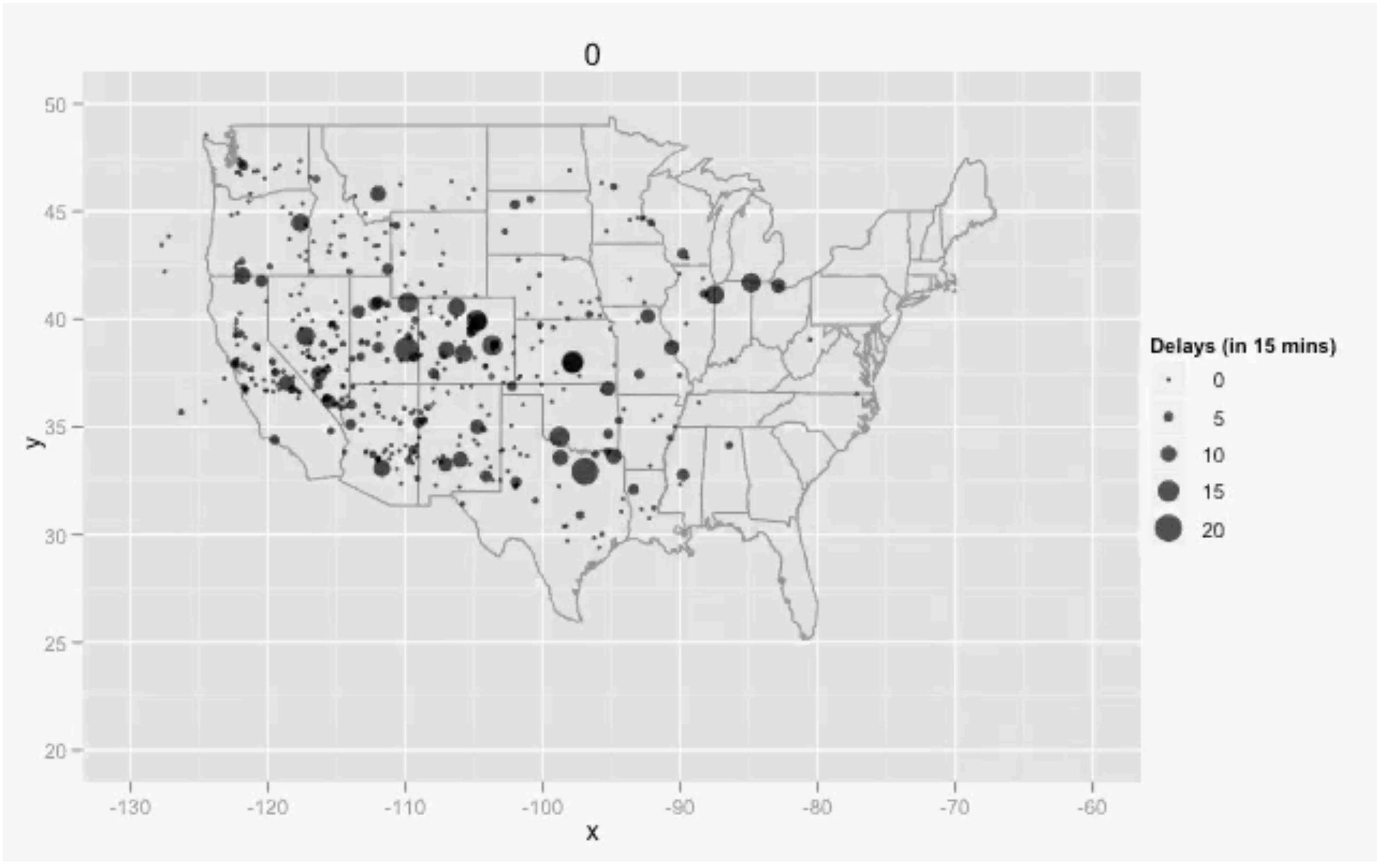
AirTran Mar 2003
46 flights,
31510 Miles
(685 mi avg.)
ATL, MCO, PHL

bankrupt in 2005

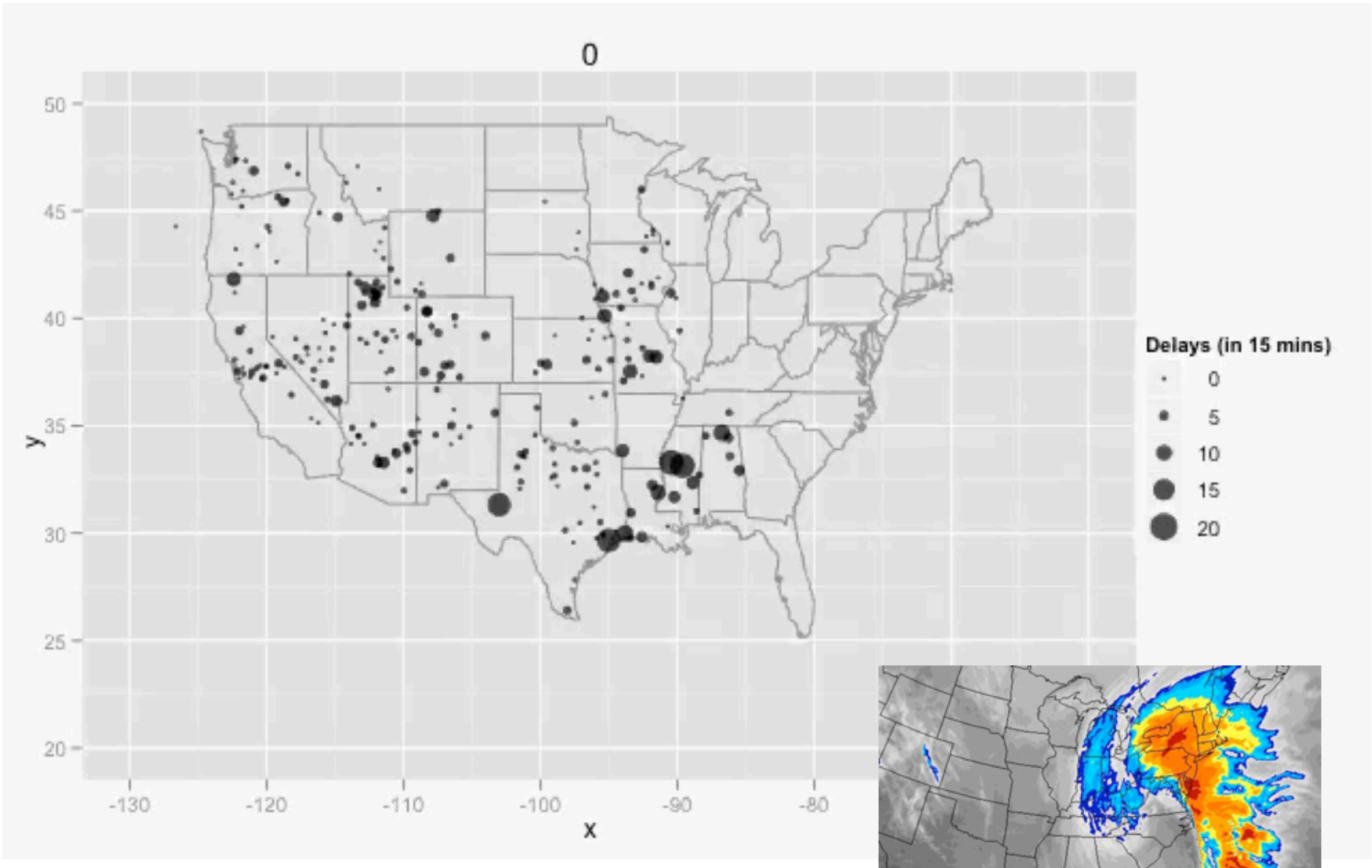


Analysis by Heike Hofmann

Can you see a difference in these two days?



Analysis by Heike Hofmann



Mar 13, 6 AM EST

Analysis by Heike Hofmann

Tidy data

- What is tidy data?
- Values in column names
- Multiple variables in one column
- Variable names in cells

Ideas, figures slides from Hadley Wickham, and tidyr publication

storms

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	10	1996-06-21

What are the variables in this dataset?

storms

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

disease counts

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

What are the variables in this dataset?

disease counts

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

pollution

city	particle size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

What are the variables in this dataset?

pollution

city	particle size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

How are these datasets similar?
How are they different?

x	y	z
1	1	2.5
2	3	4.6
1	3	1.7
4	4	7.2

2.5		1.7	
		4.6	
			7.2

How are these datasets similar?
How are they different?

x	y	z
1	1	2.5
2	3	4.6
1	3	1.7
4	4	7.2

key-value

2.5		1.7	
		4.6	
			7.2

matrix

How are these datasets similar?
How are they different?

spread

x	y	z
1	1	2.5
2	3	4.6
1	3	1.7
4	4	7.2

2.5		1.7	
		4.6	
			7.2

gather

How are these datasets similar?
How are they different?

x	y	z
W	a	2.5
X	c	4.6
W	c	1.7
Z	d	7.2

	a	b	c	d
W	2.5		1.7	
X			4.6	
Y				
Z				7.2

Tidy data = lego

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

Messy data = playmobile

Tuberculosis

http://www.flickr.com/photos/diekatrin/4299075534/

- Collected by World Health Organization
- counts of TB cases by country, year, and demographic group

What are the variables in this data?

	iso2 (chr)	year (int)	m_04 (int)	m_514 (int)	m_014 (int)	m_1524 (int)	m_2534 (int)	m_3544 (int)	m_4554 (int)	m_5564 (int)	m_65 (int)	m_u (int)
1	ZW	2003	NA	NA	133	874	3048	2228	981	367	205	NA
2	ZW	2004	NA	NA	187	833	2908	2298	1056	366	198	NA
3	ZW	2005	NA	NA	210	837	2264	1855	762	295	656	NA
4	ZW	2006	NA	NA	215	736	2391	1939	896	348	199	NA
5	ZW	2007	6	132	138	500	3693	0	716	292	153	NA
6	ZW	2008	NA	NA	127	614	0	3316	704	263	185	0

Variables not shown: f_04 (int), f_514 (int), f_014 (int), f_1524 (int), f_2534 (int), f_3544 (int), f_4554 (int), f_5564 (int), f_65 (int), f_u (int)

gather make key-value pair

separate “f_014” becomes “f” “014”

arrange sort by year, sex, age

```
  country  year    sex    age cases
    (chr) (int) (chr) (chr) (int)
1      AD  1996      f    014     0
2      AD  1996      f   1524     1
3      AD  1996      f   2534     1
4      AD  1996      f   3544     0
5      AD  1996      f   4554     0
6      AD  1996      f   5564     1
```

Your turn

Variables in cells

Melbourne weather records from GHCN

What are the variables?

	V1	V2	V3	V4	V5	V9	V13	V17	V21	V25	V29	V33	V37	V41	V45	V49	V53
1	ASN00086282	1970	7	TMAX	141	124	113	123	148	149	139	153	123	108	119	112	126
2	ASN00086282	1970	7	TMIN	80	63	36	57	69	47	84	78	49	42	48	56	51
3	ASN00086282	1970	7	PRCP	3	30	0	0	36	3	0	0	10	23	3	0	5
4	ASN00086282	1970	8	TMAX	145	128	150	122	109	112	116	142	166	127	117	127	159
5	ASN00086282	1970	8	TMIN	50	61	75	67	41	51	48	-7	56	62	47	33	67
6	ASN00086282	1970	8	PRCP	0	66	0	53	13	3	8	0	0	0	3	5	0

gather make key-value pair for days

spread separate columns for tmin, max, prep

new vars compute new variables

	stn	year	month	day	tmin	tmax	t_range	prcp	date
1	ASN00086282	1970	7	1	8.0	14.1	6.1	0.3	1970-07-01
2	ASN00086282	1970	7	2	6.3	12.4	6.1	3.0	1970-07-02
3	ASN00086282	1970	7	3	3.6	11.3	7.7	0.0	1970-07-03
4	ASN00086282	1970	7	4	5.7	12.3	6.6	0.0	1970-07-04
5	ASN00086282	1970	7	5	6.9	14.8	7.9	3.6	1970-07-05
6	ASN00086282	1970	7	6	4.7	14.9	10.2	0.3	1970-07-06

TIDY DATA
can be summarised,
plotted and analysed
EFFICIENTLY

Tuberculosis

- Tuberculosis incidence is easily counted by country, by gender, by age, by year
- Investigate questions such as “Is the incidence increasing?”, “Is it more likely to occur in children?”, “Is the prevalence higher for girls in some countries?”

BUT WHAT INFORMATION IS MISSING???

Tuberculosis

- We need populations for the different countries and years in order to compute rates, and then comparisons can be made
- Ok, got it and computed rates, cases/population

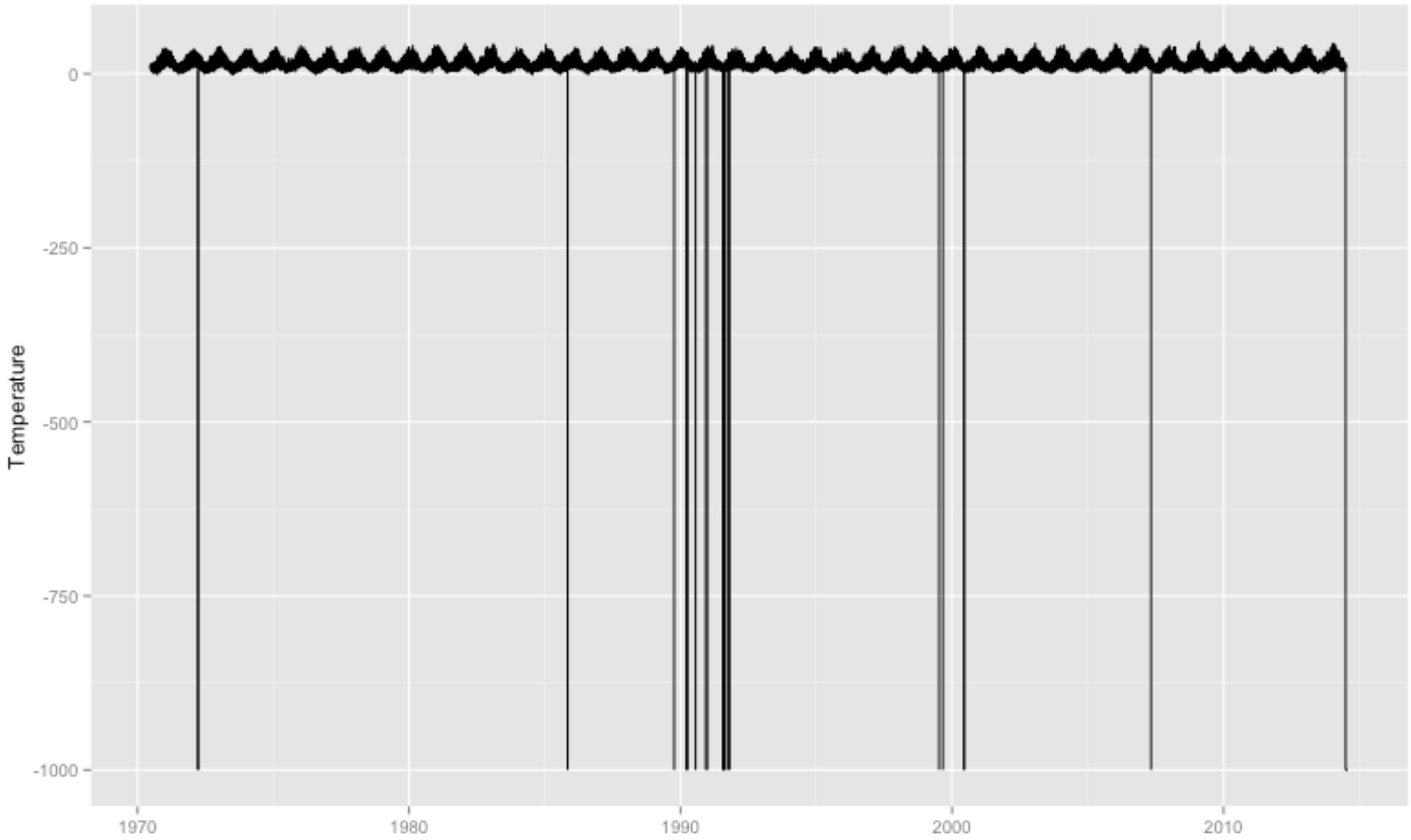
HIGHEST

	country (chr)	year (int)	sex (chr)	age (chr)	cases (int)	population (int)	rate (dbl)
1	TK	2008	m	4554	9	59	0.15254237
2	TK	2008	m	2534	10	82	0.12195122
3	TK	2008	f	1524	13	113	0.11504425
4	TK	2008	f	65	4	35	0.11428571
5	TK	2008	m	5564	4	35	0.11428571
6	TK	2008	f	4554	6	56	0.10714286
7	TK	2008	f	5564	3	35	0.08571429
8	TK	2008	m	1524	10	122	0.08196721
9	TK	2008	f	2534	6	77	0.07792208
10	TK	2008	f	3544	5	75	0.06666667
..	...	...	...	...	...	...	...

What country is TK?

Temperatures

- With tidy data it is easy to examine temperature ranges, and precipitation over time.
- Investigate questions such as “Is temperature getting more extreme?”, “Is precipitation declining?”



Is this what you expected?

Missing data

- Some values are not collected
- How is this coded?
- What are the effects?
- Imputing missings

Airline traffic

RR_TIME	ARR_DELAY	CRS_EL_TIME	ACT_EL_TIME	AIR_TIME	DIST
0239		220.00			1514.00
0055	-15.00	265.00	255.00	235.00	1721.00
1835	5.00	127.00	118.00	91.00	622.00
1620	-18.00	208.00	191.00	172.00	1491.00
1004	-18.00	256.00	240.00	214.00	1605.00

What is missing?

Tuberculosis

	iso2 (chr)	year (int)	m_04 (int)	m_514 (int)	m_014 (int)	m_1524 (int)	m_2534 (int)	m_3544 (int)	m_4554 (int)	m_5564 (int)	m_65 (int)	m_u (int)
1	ZW	2003	NA	NA	133	874	3048	2228	981	367	205	NA
2	ZW	2004	NA	NA	187	833	2908	2298	1056	366	198	NA
3	ZW	2005	NA	NA	210	837	2264	1855	762	295	656	NA
4	ZW	2006	NA	NA	215	736	2391	1939	896	348	199	NA
5	ZW	2007	6	132	138	500	3693	0	716	292	153	NA
6	ZW	2008	NA	NA	127	614	0	3316	704	263	185	0

Variables not shown: f_04 (int), f_514 (int), f_014 (int), f_1524 (int), f_2534 (int), f_3544 (int), f_4554 (int), f_5564 (int), f_65 (int), f_u (int)

What is missing?

Melbourne weather

What is missing?

	stn	year	month	day	tmin	tmax	t_range	prcp	date
16394	ASN00086282	2014	7	26	-999.9	-999.9	0	-999.9	2014-07-26
16395	ASN00086282	2014	7	27	-999.9	-999.9	0	-999.9	2014-07-27
16396	ASN00086282	2014	7	28	-999.9	-999.9	0	-999.9	2014-07-28
16397	ASN00086282	2014	7	29	-999.9	-999.9	0	-999.9	2014-07-29
16398	ASN00086282	2014	7	30	-999.9	-999.9	0	-999.9	2014-07-30
16399	ASN00086282	2014	7	31	-999.9	-999.9	0	-999.9	2014-07-31

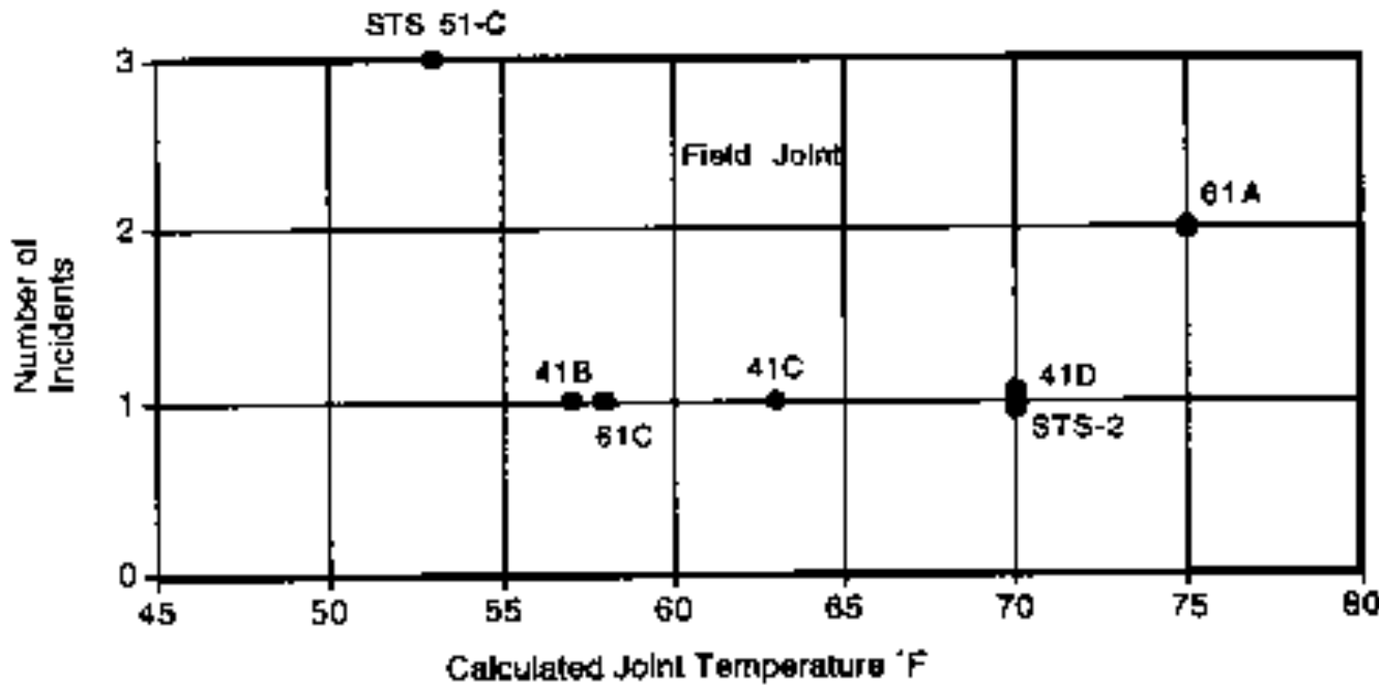
http://bit.ly/wrangling3

Your turn

What ways have you seen missing coded?

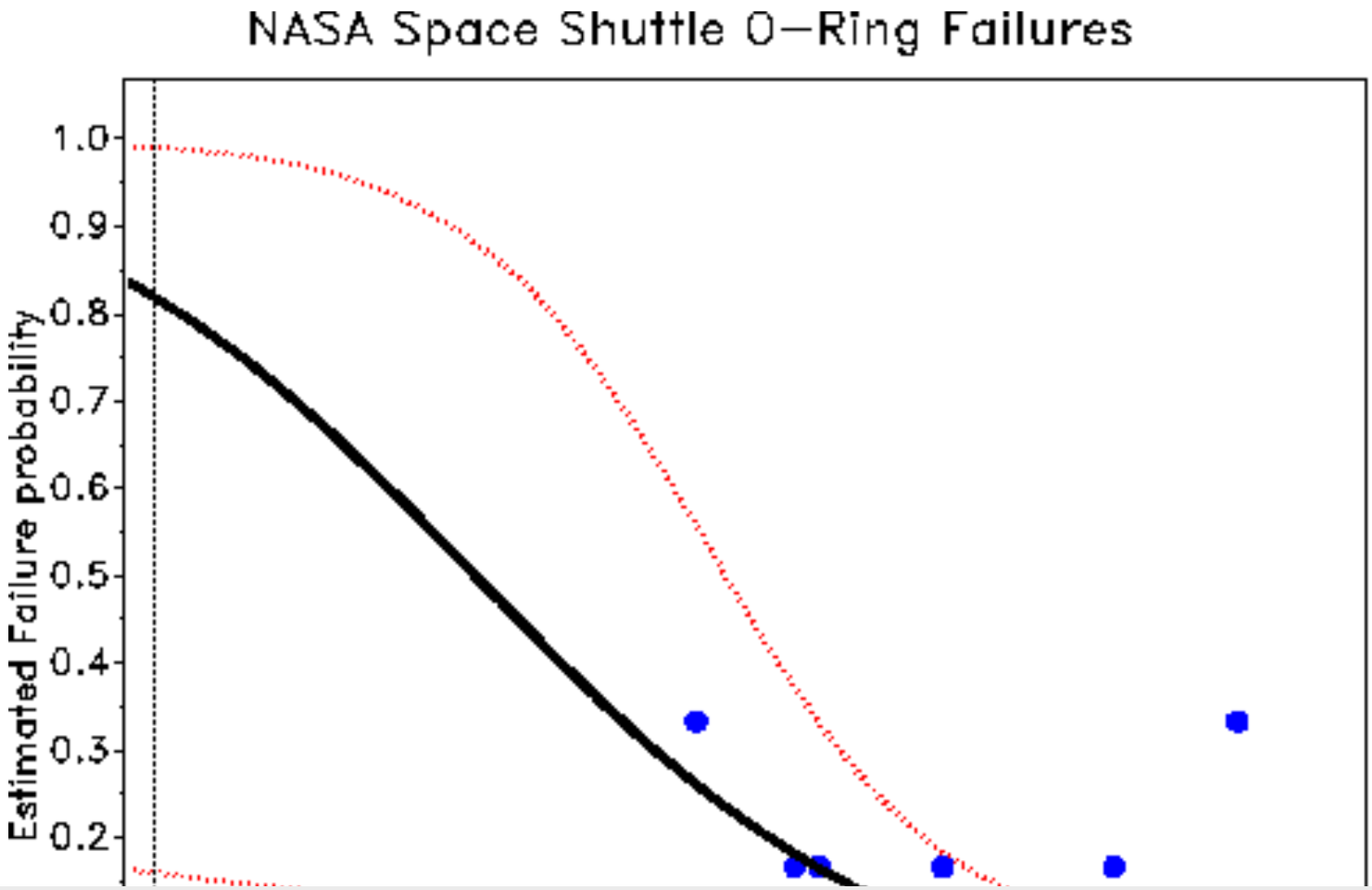
Challenger disaster

- Subsequent investigation determined that the cause was failure of the O-ring seals used to isolate the fuel supply from burning gases.
- NASA staff had analysed the data on the relation between ambient temperature and number of O-ring failures (out of 6), **but they had excluded observations where no O-rings failed**, believing that they were uninformative. Unfortunately, those observations had occurred when the launch temperature was relatively warm (65-80 degF).



This figure (scanned badly from Wainer, 1995) shows a graph accompanying the Report of the Presidential Commission on the Space Shuttle Challenger Accident, 1986 (vol 1, p. 145) in the aftermath of the disaster.

Missings are IMPORTANT



Re-analysis of the O-ring data involved fitting a logistic regression model. This provides a predicted extrapolation (black curve) of the **probability of failure to the low (31 degF) temperature** at the time of the launch and confidence bands on that extrapolation (red curves).

Tappin, L. (1994). "Analyzing data relating to the Challenger disaster". *Mathematics Teacher*, 87, 423-426

Traps

- Software often drops missing values without informing you, producing quite probably misleading results and biased estimates
- Many modeling algorithms require complete data

Handling missings

- Summarise
  - Proportion of missing in each column
  - Proportion of missing in each row
  - Overall proportion of missing values
  - Stratified by categories in the data
- Explore distribution of missing vs not missing - eg. do missings occur more often if humidity is high?
- Impute & check

Numeric Summary for Missing Values

Missing:

3.01% of the numbers

37.5% of variables

23.23% of samples

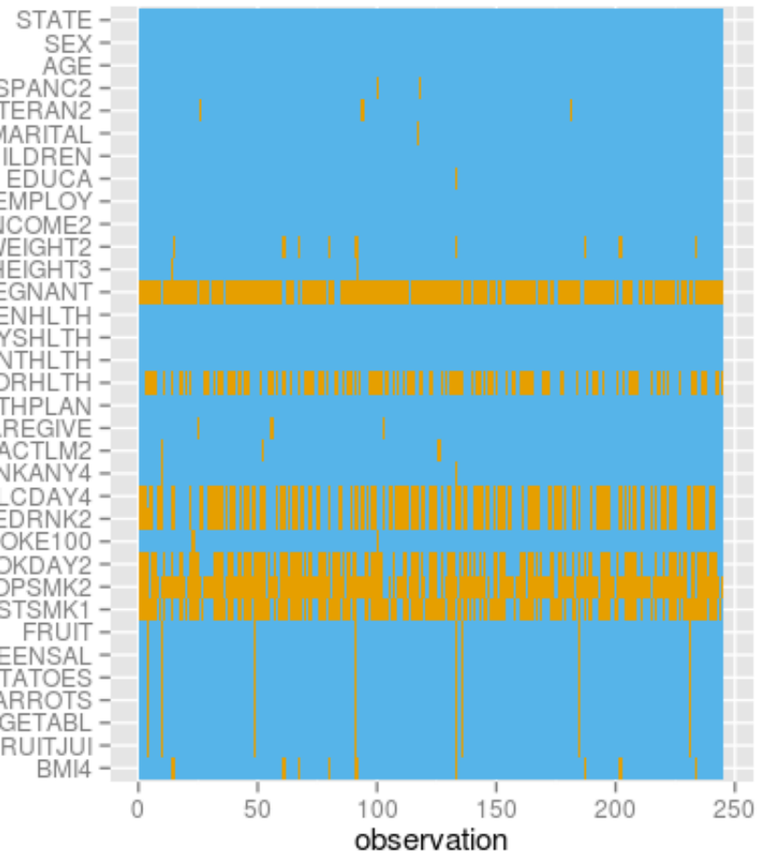
No_of_miss_by_case	No_of_Case	Percent
0	565	76.8
1	167	22.7
2	2	0.3
3	2	0.3
4	0	0
5	0	0
6	0	0
7	0	0
8	0	0

Missing Values

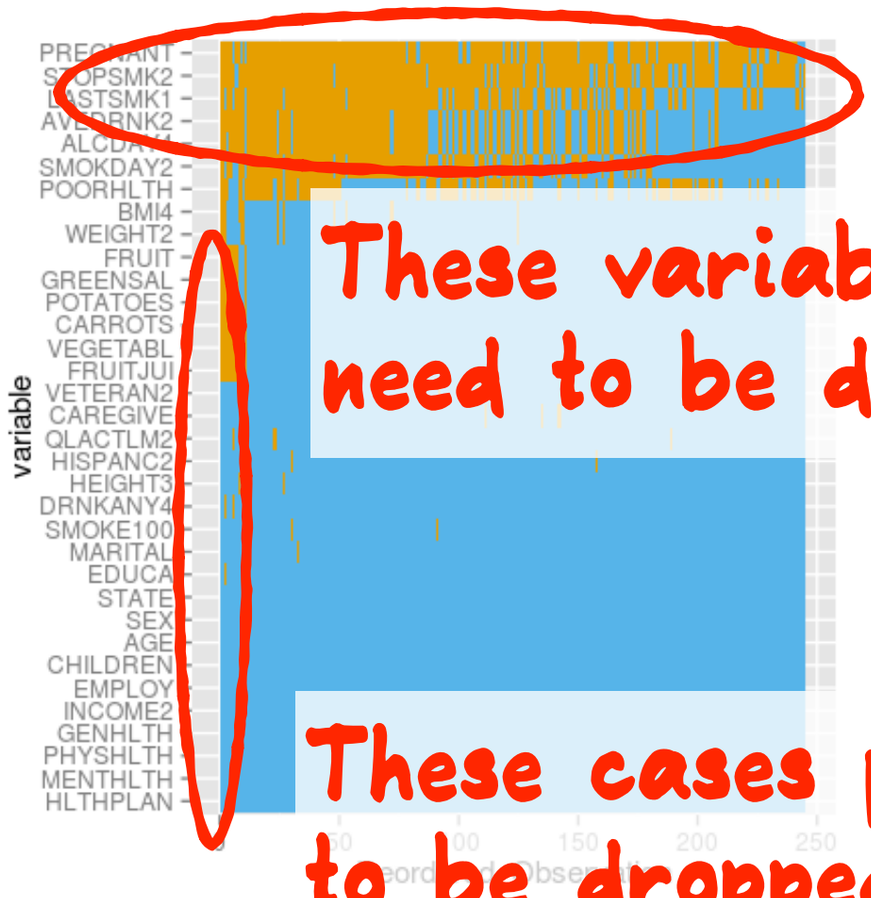
Summary | Help | Settings

ID	Variables	Class	NApct
1	year	factor	0
2	latitude	factor	0
3	longitude	factor	0
4	sea.surface.temp	numeric	0.004
5	air.temp	numeric	0.11
6	humidity	numeric	0.126
7	uwind	numeric	0
8	vwind	numeric	0

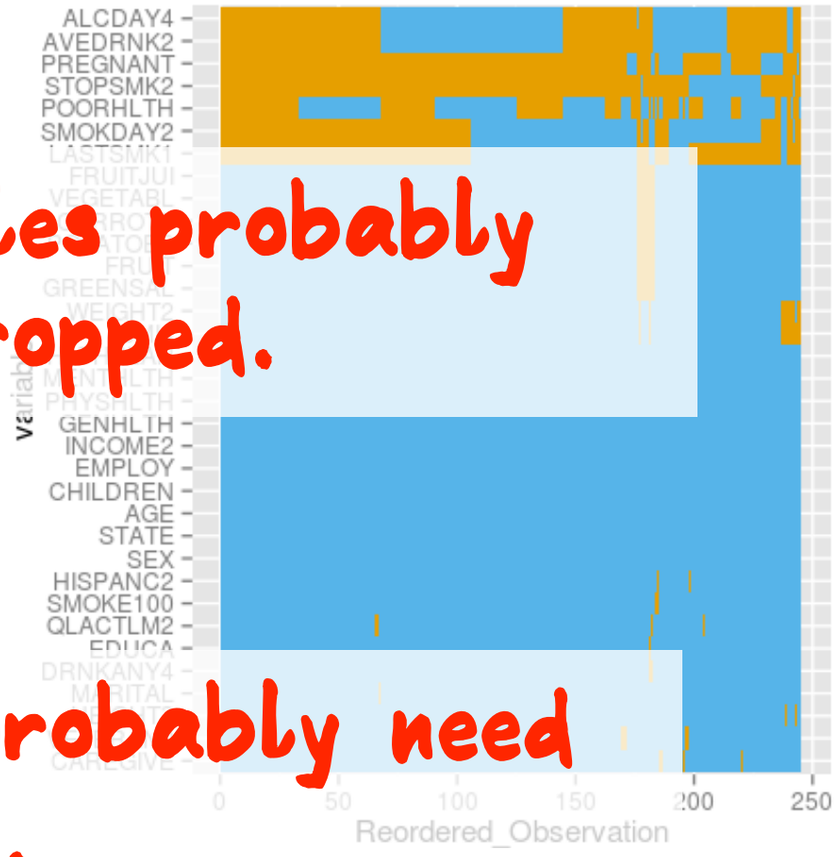
Missingness map of a data set



Original order



Ordered by missingness



Clustered by missingness

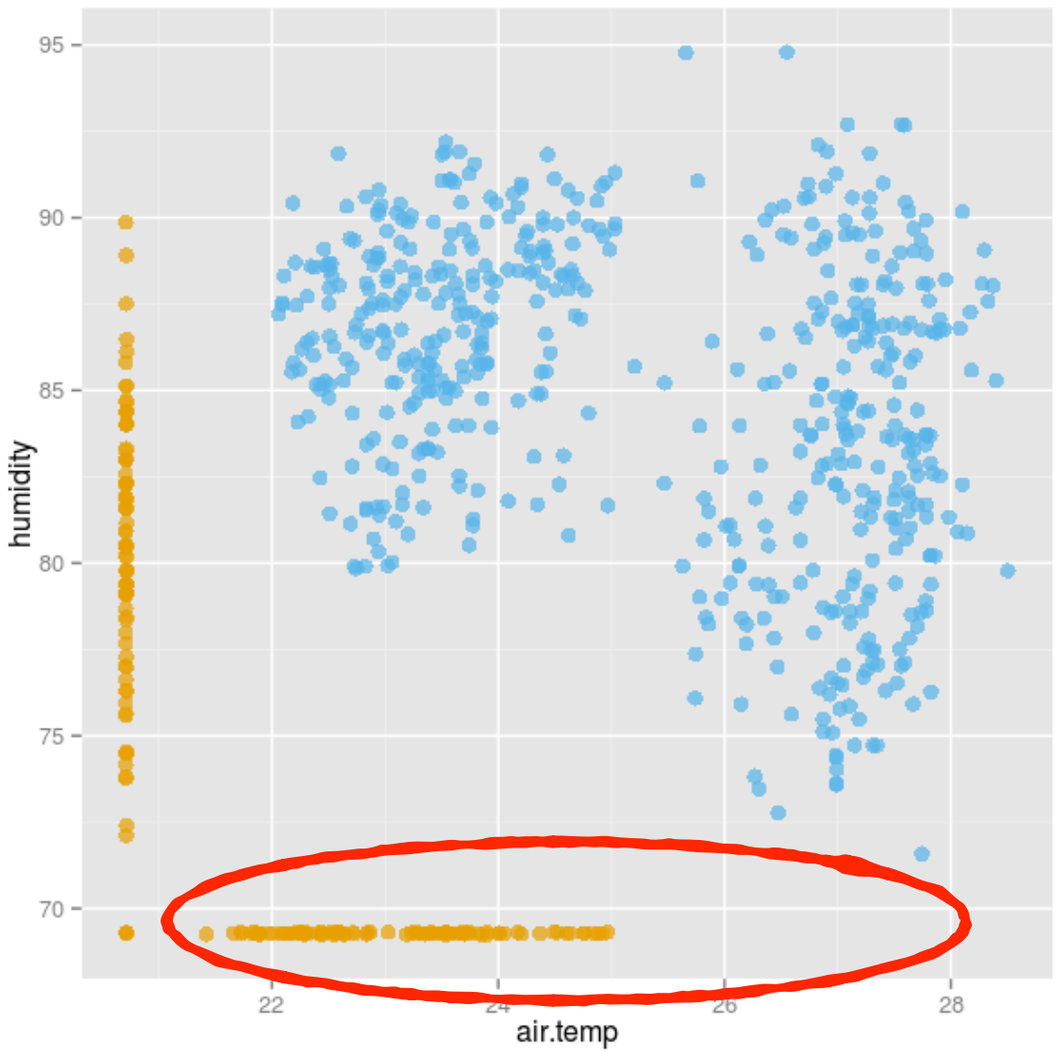
These variables probably need to be dropped.

These cases probably need to be dropped.

What is do we learn?

Missing dependencies

- Plot the variables and calculate summaries for missing or not missing on another variable

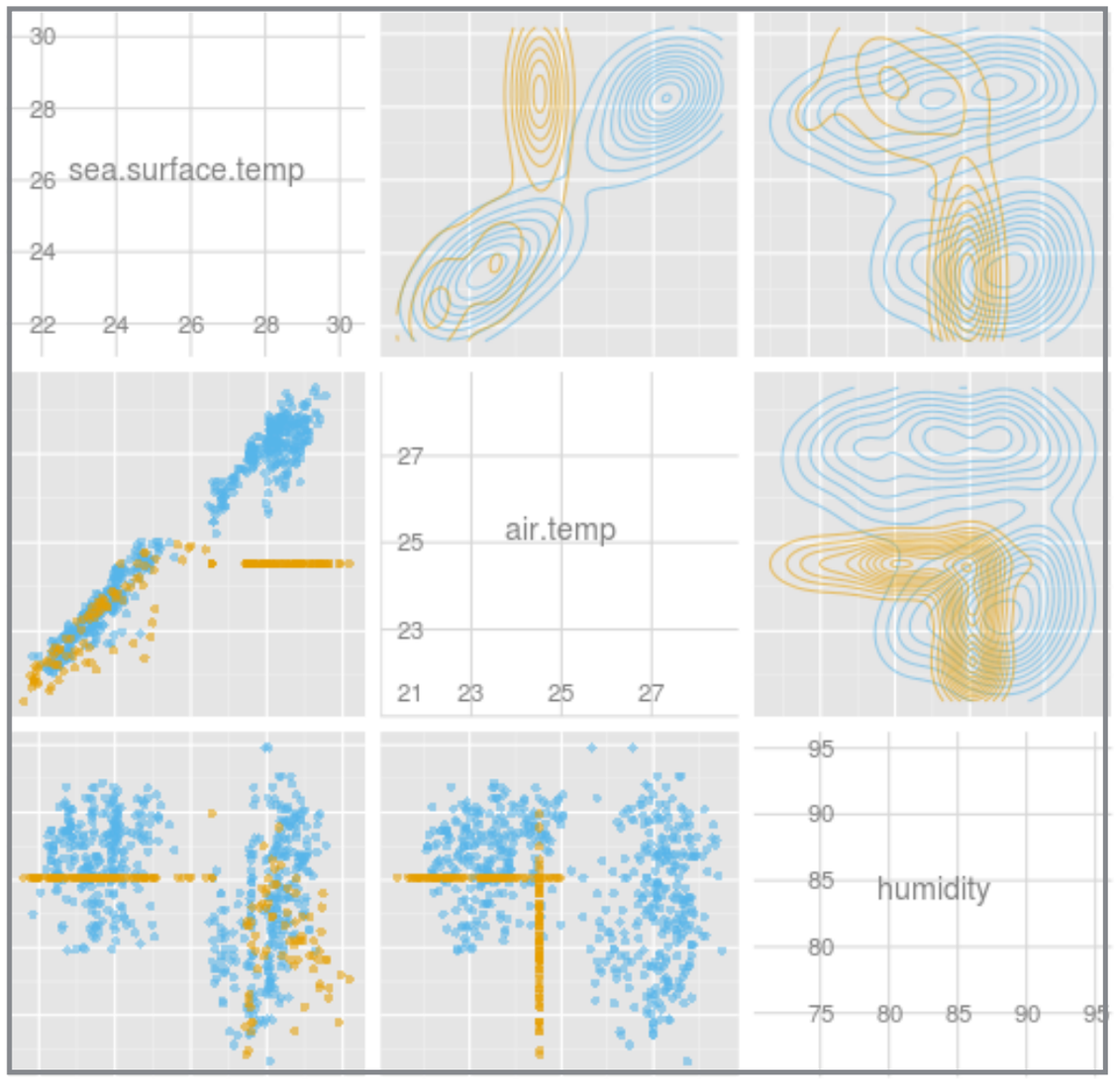


Missing values on HUMIDITY (plotted on y-axis), but not missing on air temperature (plotted on x-axis)

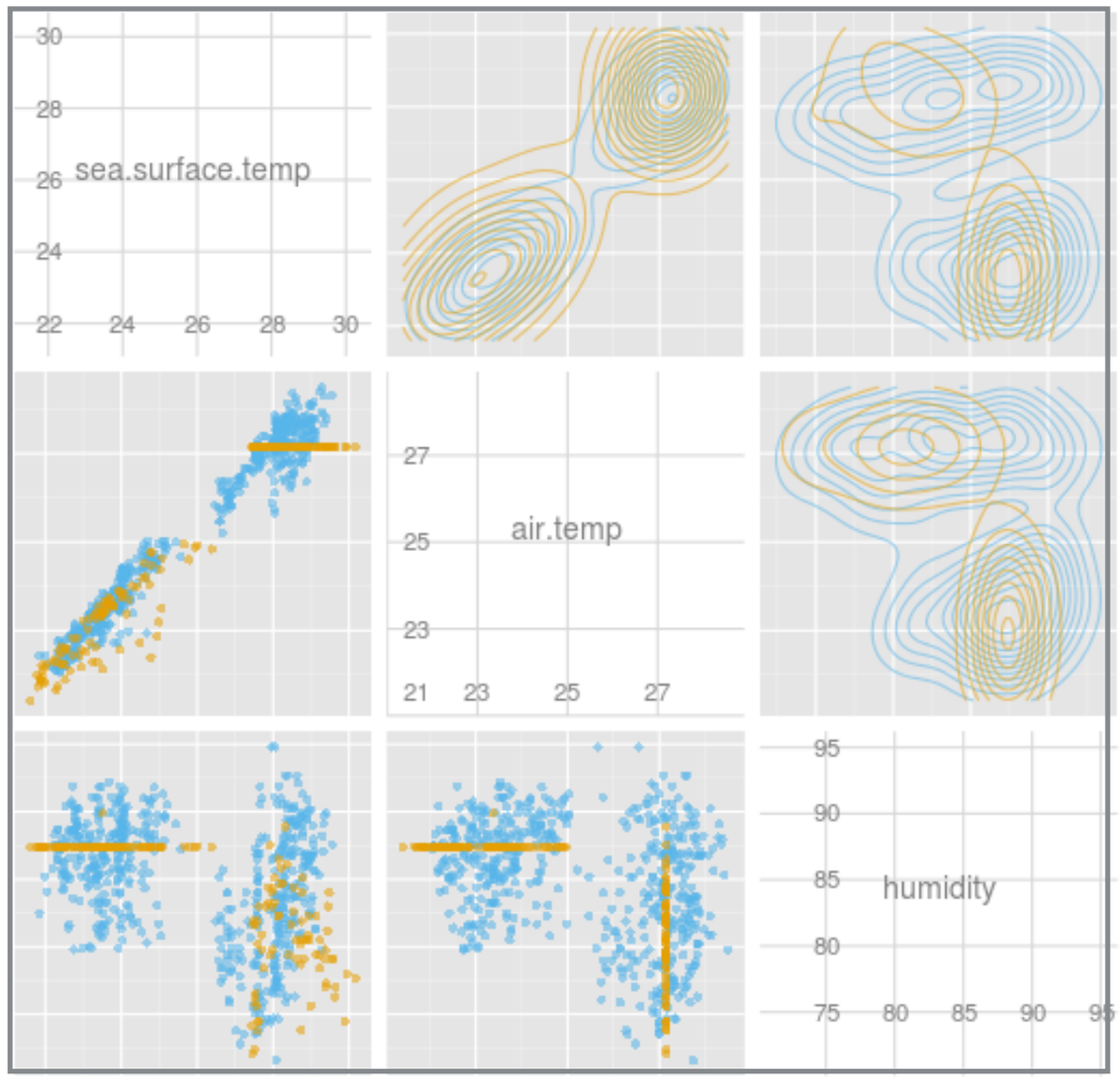
Imputation

- Mean/median of complete cases, simple sometimes reasonable
- Average the values of the nearest neighbors, using the complete variables
- Use a model, like regression to predict the missings, based on the complete cases
- Simulate from a probability model, like a normal distribution using the sample statistics of complete cases as parameters

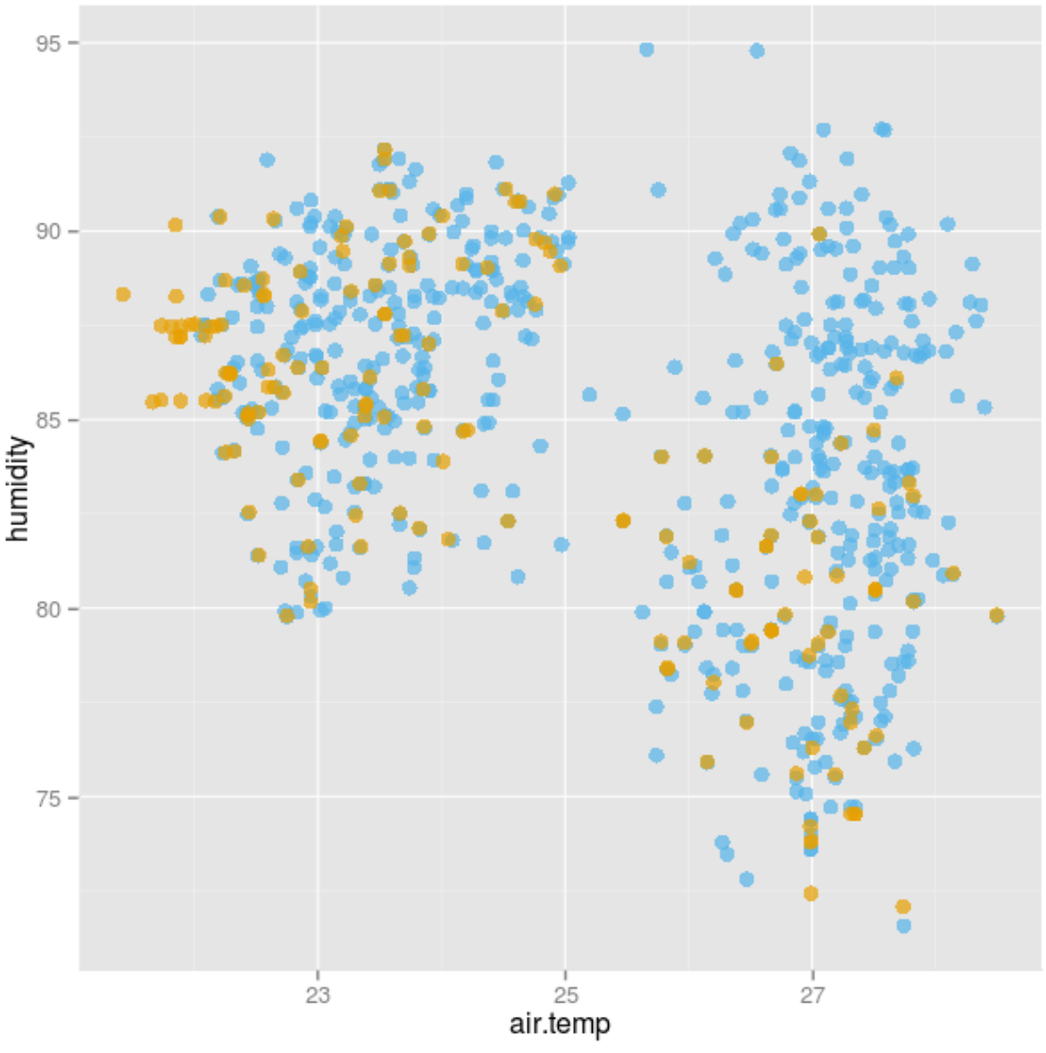
Missings imputed using the **overall** median of complete cases



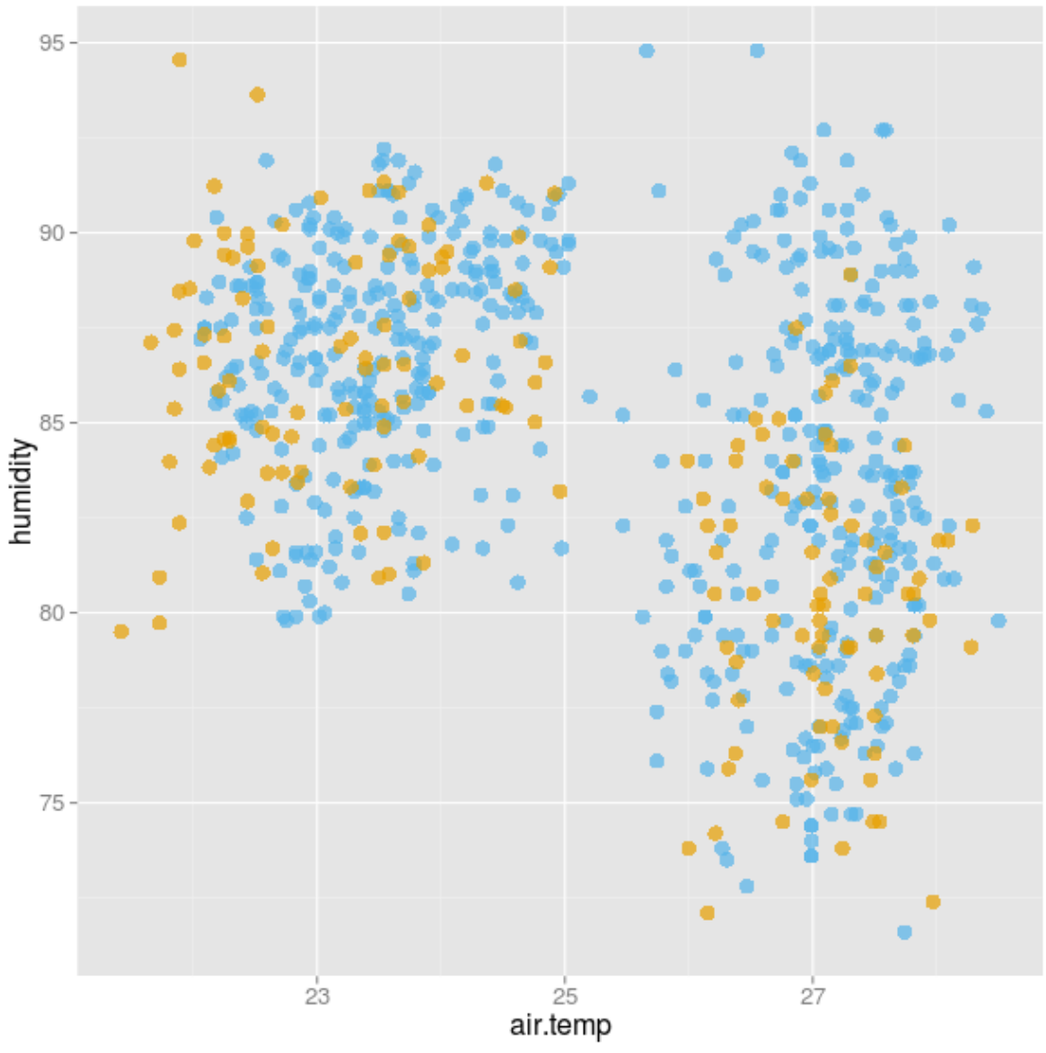
Missings imputed using the median of complete cases **by cluster**

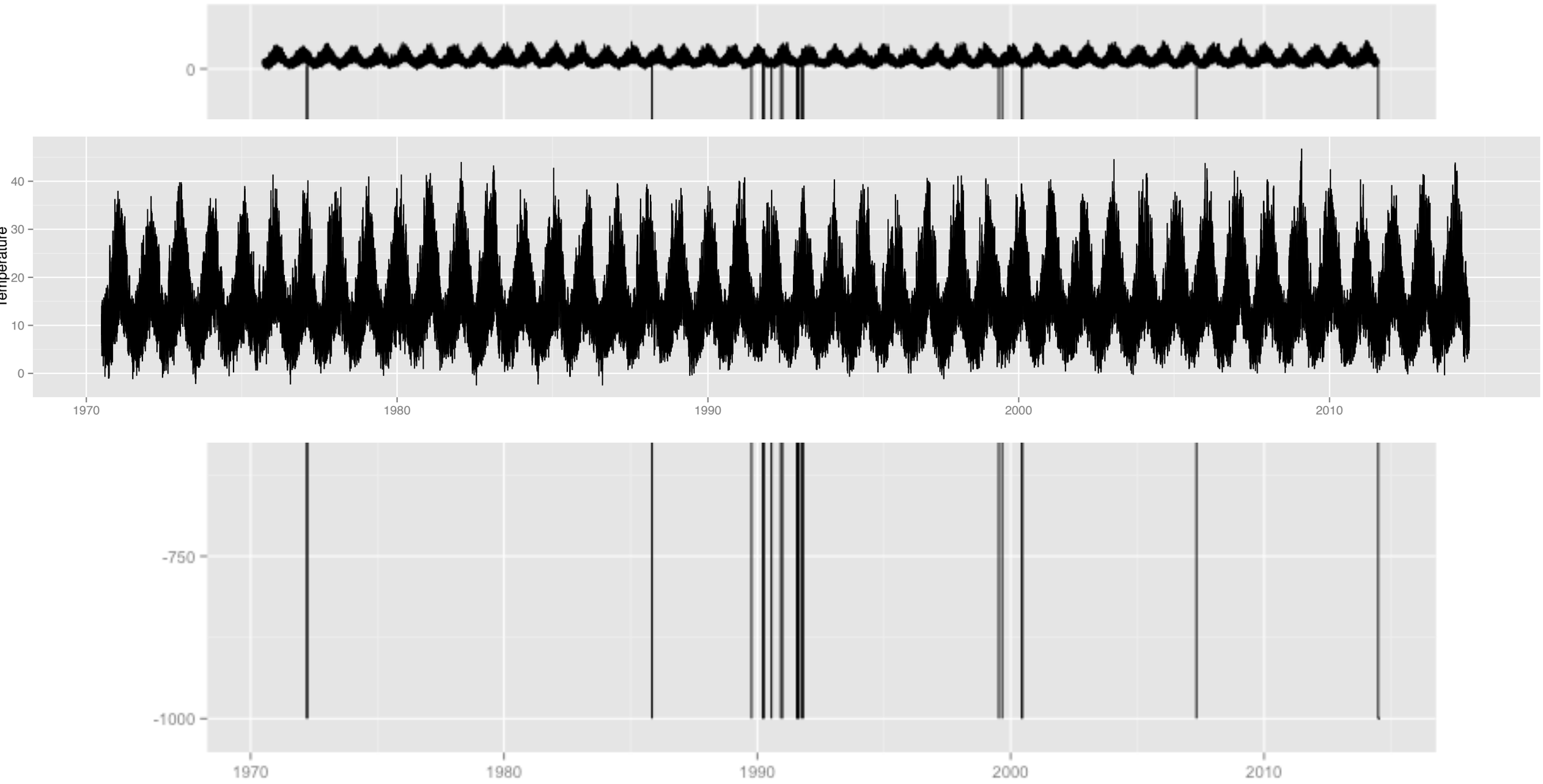


Missings imputed using **five nearest neighbours**



Missings imputed using **simulation from a multivariate normal model**





Temperature data re-plotted after -999.9's handled

What else?

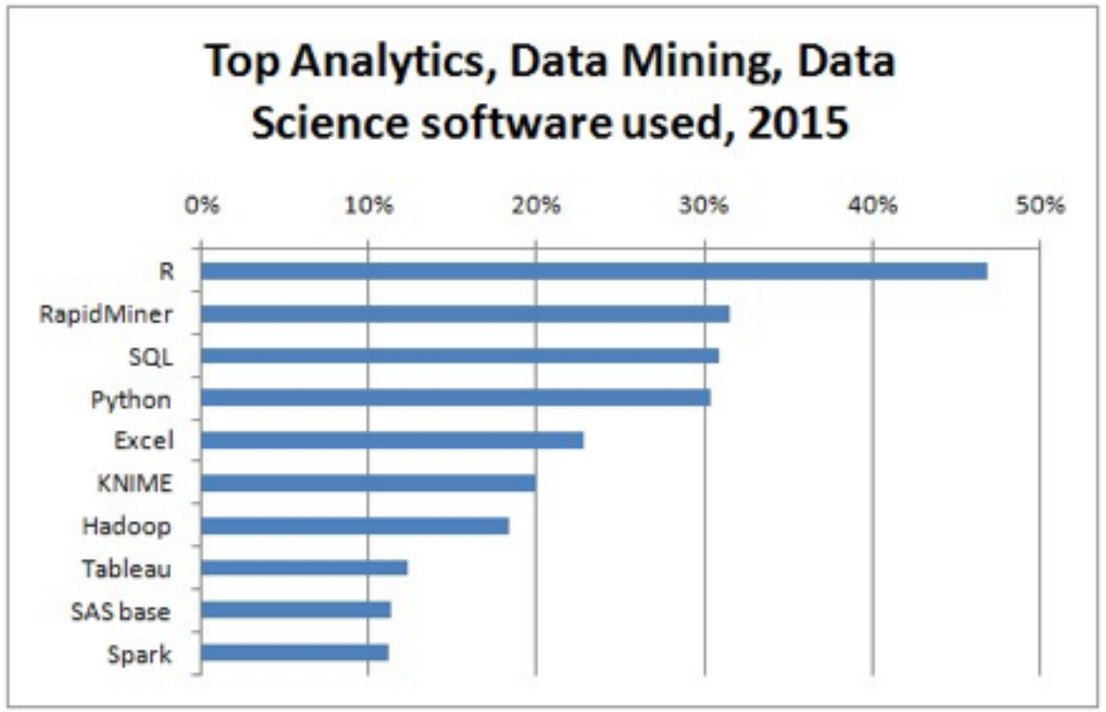
- Web scraping skills are really useful
- Reproducible reports can help in the long term
- Use the computer to understand what might have happened by chance - simulation, sampling, permutation are really useful

"With the wide availability of computer packages and graphics nowadays there is no excuse for ducking the labour of this preliminary phase, and it may save some red faces later."
Crowder and Hand (1990)

2015 Best Data Tools

http://www.kdnuggets.com/polls/2015/analytics-data-mining-data-science-software-used.html

1. R, 46.9% share (38.5% in 2014)
2. RapidMiner, 31.5% (44.2% in 2014)
3. SQL, 30.9% (25.3% in 2014)
4. Python, 30.3% (19.5% in 2014)
5. Excel, 22.9% (25.8% in 2014)
6. KNIME, 20.0% (15.0% in 2014)
7. Hadoop, 18.4% (12.7% in 2014)
8. Tableau, 12.4% (9.1% in 2014)
9. SAS, 11.3 (10.9% in 2014)



R packages for data wrangling

- devtools
- ggplot2
- readr
- tidyr
- dplyr
- rvest
- stringr
- lubridate
- knitr

- jsonlite
- RMySQL
- XML
- shiny
- ggmap

http://www.computerworld.com/article/2921176/business-intelligence/great-r-packages-for-data-import-wrangling-visualization.html


"Time to play in the sandbox is really helpful for developing skills. Data competitions like available at kaggle encourage learning about working with data"
Di Cook, Data Wrangling for Managers, 2015