A beginner's guide to using SQL with R: database usage for fun and profit

Nicholas J. Horton

Amherst College, Amherst, MA, USA

October 8, 2013

nhorton@amherst.edu

Acknowledgements

- joint work with Ben Baumer (Smith College) and Hadley Wickham (Rice/RStudio)
- supported by NSF grant 0920350 (building a community around modeling, statistics, computation and calculus)
- more information at `http://www.mosaic-web.org`

Cautionary Note

ACM White Paper on Data Science (`www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf`)

> *The promise of data-driven decision-making is now being recognized broadly, and there is growing enthusiasm for the notion of "Big Data." (first line)*

Cautionary Note

ACM White Paper on Data Science (`www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf`)

> *The promise of data-driven decision-making is now being recognized broadly, and there is growing enthusiasm for the notion of "Big Data." (first line)*
>
> *Methods for querying and mining Big Data are fundamentally different from traditional statistical analysis on small samples. (first mention of statistics, page 7)*

Cautionary Note

ACM White Paper on Data Science (`www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf`)

> *The promise of data-driven decision-making is now being recognized broadly, and there is growing enthusiasm for the notion of "Big Data." (first line)*
>
> *Methods for querying and mining Big Data are fundamentally different from traditional statistical analysis on small samples. (first mention of statistics, page 7)*

Do statisticians just provide old-school tools for use by the new breed of data scientists?

Cautionary Note (cont.)

- Cobb argued (TISE, 2007) that our courses teach techniques developed by pre-computer-era statisticians as a way to address their lack of computational power
- Do our students see the potential and exciting use of statistics in our classes? (Gould, ISR, 2010)
- How do we respond to these external and internal challenges?

Prelude

- "New Frontier in statistical thinking" (Chamandy, Google @ JSM 2013)
- teaching precursors to big data/foundations of data science as an intellectually coherent theme (Pruim, Calvin College)
- growing importance of computing and ability to "think with data" (Lambert, Google)
- key capacities in statistical computing (Nolan and Temple Lang, TAS 2010)
- "Statistics and the modern student" (Gould, ISR 2010)

Prelude (cont.)

How to accomplish this?

- start in the first course
- build on capacities in the second course
- develop more opportunities for students to apply their knowledge in practice (internships, collaborative research, teaching assistants)
- new courses focused on "Data Science"
- "Data Expo" and "Data Fest" opportunities
- today's goal: talk about what can be done early...

Data Expo 2009

Ask students: have you ever been stuck in an airport because your flight was delayed or cancelled and wondered if you could have predicted it if you'd had more data? (Wickham, JCGS, 2011)

Data Expo 2009

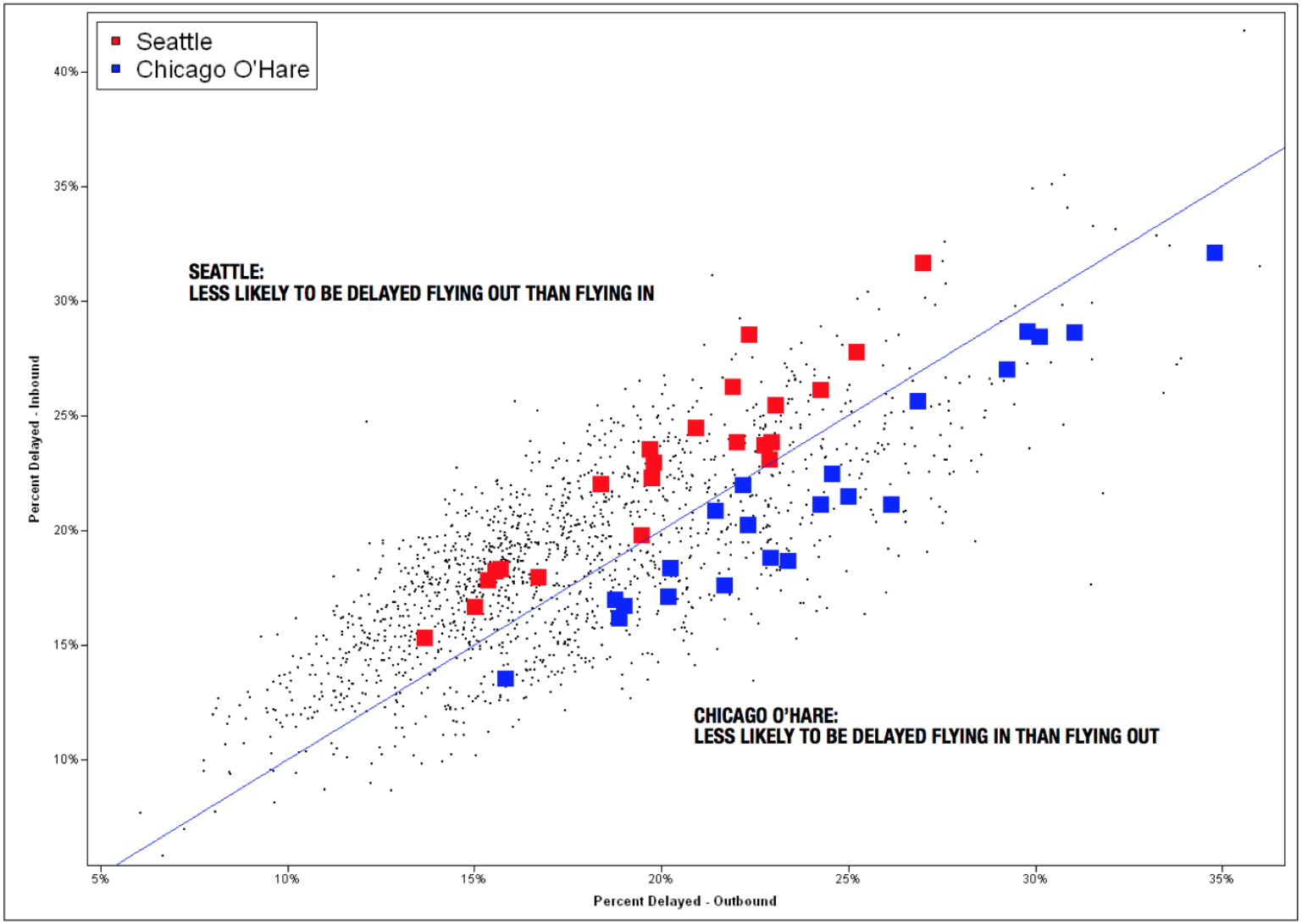
- dataset of flight arrival and departure details for all commercial flights within the USA, from October 1987 to April 2008 (but we now have through the end of 2012!)
- large dataset: more than 150 million records
- aim: provide a graphical summary of important features of the data set
- winners presented at the JSM in 2009; details at `http://stat-computing.org/dataexpo/2009`

Airline Delays Codebook (abridged)

Year	$1987, 1998, \ldots, 2012$
Month	1 through 12
DayofMonth	1 through 31
DayOfWeek	1=Monday, 7=Sunday
DepTime	departure time
UniqueCarrier	OH = Comair, DL = Delta, etc.
TailNum	plane tail number
ArrDelay	arrival delay, in minutes
Origin	BDL, BOS, MSP, PHX, SFO, etc.
Dest	

Full details at
`http://www.transtats.bts.gov/Fields.asp?Table_ID=236`

Sampling of the Data Expo 2009 winners



Sampling of the Data Expo 2009 winners

CAN WE SEE WHAT IS NOT THERE?

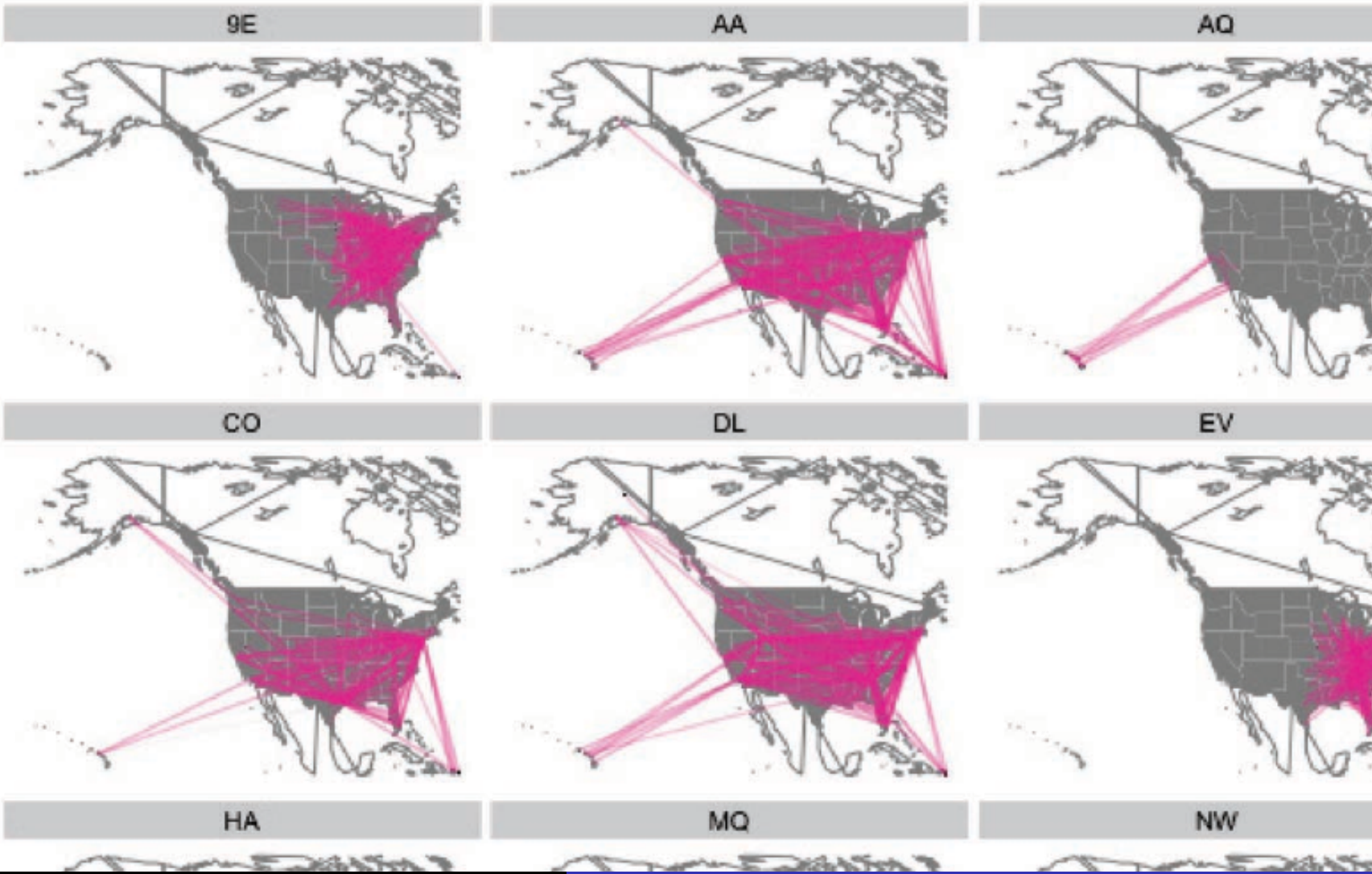
Planes have, for reasons such as maintenance, weather, or schedule fly empty between airports as so-called *Ghosts*. By tracking individual planes, we reveal their paths, including situations, where a plane lands in a different airport than where it takes off later, i.e. a ghost:

Example: US Airways Aircraft N-881 - Ghostflight from PIT to RIC (222 miles)

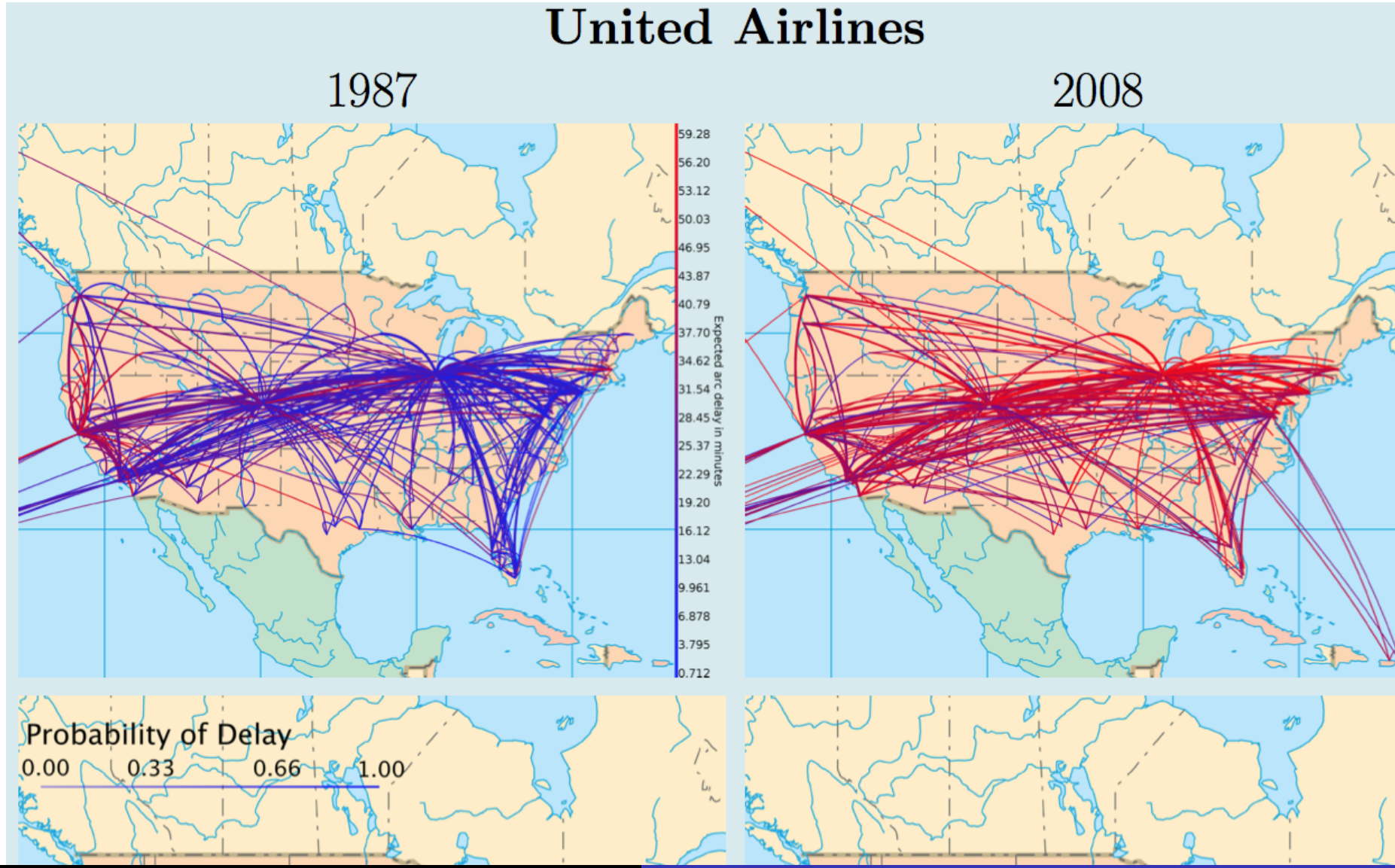
Year	Month	Day	DepTime	ArrTime	Origin	Dest	Diverted
1995	3	8	1102	1256	PIT	CVG	0
1995	3	8	1311	NA	CVG	PIT	1
1995	3	8	1913	2050	RIC	PIT	0
1995	3	8	2134	2300	PIT	MSY	0

Ghost Flight Totals: over 1 million flights since 1995, with an average dis-

Sampling of the Data Expo 2009 winners



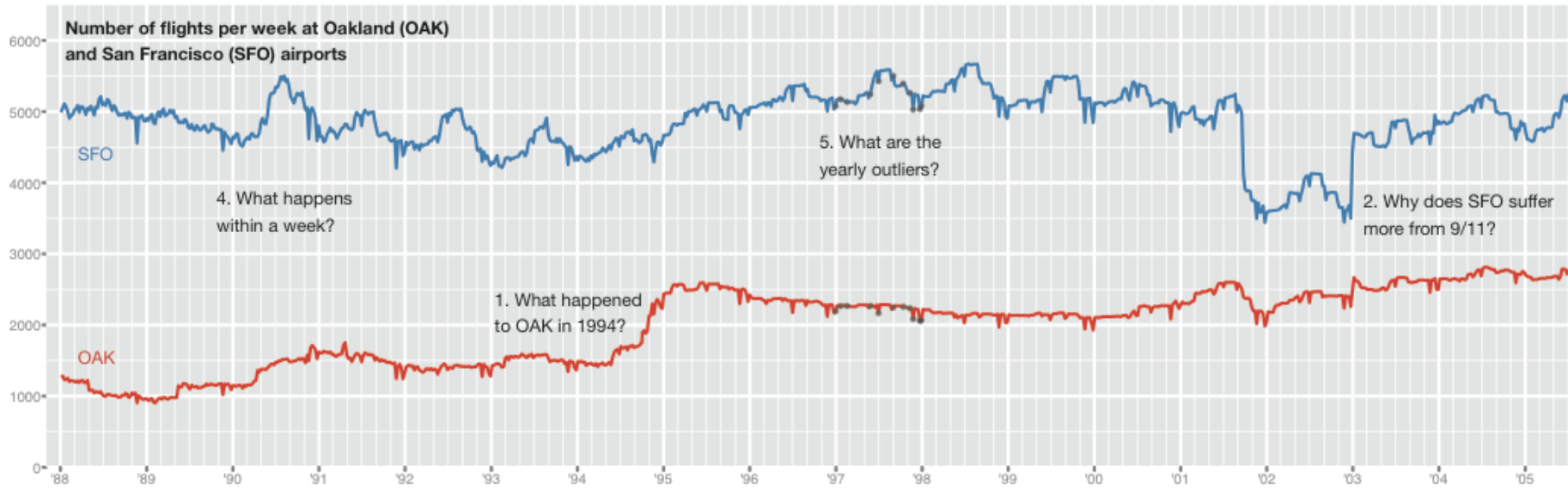
Sampling of the Data Expo 2009 winners



Sampling of the Data Expo 2009 winners

A Tale of Two Airports

AN EXPLORATION OF FLIGHT TRAFFIC AT OAK AND SFO



Why didn't I participate?

- didn't have experience using databases
- lack of time to learn new (?) technologies
- need to combine multiple tools

Goal

- make a complex and interesting dataset accessible to students in introductory statistics
- facilitate use of database technology for instructors without training in this area
- demonstrate how to use SQL (using MySql and/or SQLite) to achieve this goal
- help faculty energize the next generation of data scientists

Background on databases and SQL

- relational databases (invented in 1970)
- like electronic filing cabinets to organize masses of data (terabytes)
- fast and efficient
- useful reference: *Learning MySQL*, O'Reilly 2007

Client and server model

- server: manages data
- client: ask server to do things
- use R as the client (using an add-on package such as `RMySQL` or `RSQLite`)

SQL

- Structured Query Language
- special purpose programming language for managing data
- developed in early 1970's
- standardized (multiple times)
- most common operation is query (using `SELECT`)

SQLite

advantage: free, quick, dirty, simple (runs locally)

disadvantage: not as robust, fast, or flexible than other free alternatives such as MySQL (which run remotely)

For personal use, or to get started SQLite is ideal.
For a class, I'd recommend MySQL. (We'll be using this today, using a database housed at Smith College).

Creating the airline delays database

1. download and install SQLite from `sqlite.org`
2. download the data (1.6gb compressed, 12gb uncompressed)
3. create a table with fields that match the csv files
4. load the data with the `.import` directive
5. add indices (to speed up access to the data, takes some time)
6. install and load the `RSQLite` package
7. establish a connection (using `dbConnect()`)
8. start to make selections (which will be returned as data frames) using the `dbGetQuery()` function

Hadley Wickham's idioms for dealing with big(ger) data

select: subset variables

filter: subset rows

mutate: add new columns

summarise: reduce to a single row

group-by: aggregate

Hadley Wickham, `bit.ly/bigrdata4`

Accessing the database

```
# establish the connection
require(RMySQL)
con = dbConnect(MySQL(), host="rucker.smith.edu",
  dbname="airlines")
# count the number of records in the database
ds = dbGetQuery(con, "SELECT COUNT(*) FROM ontime")

  COUNT(*)
1 1.49e+08
```

Accessing the database

```
# count flights by airplanes
> ds = dbGetQuery(con, "SELECT COUNT(*),
     tailnum FROM ontime GROUP BY tailnum")
> dim(ds)
[1] 15589     2
> sorted = ds[order(ds[,1], decreasing=TRUE),]
> head(sorted)
      COUNT(*)      tailnum
1     37888947
15587   572311       UNKNOW
15589   170932 \xe4NKNO\xe6
6831     35879       N477HA
6919     35844       N486HA
```

Tail Number N477HA is a 2001 BOEING 717-200

Share Tweet Share Like 0 Send

Limited Trial – Normally for Reg

Aircraft Information

SERIAL NUMBER
55122

AIRCRAFT TYPE
Fixed Wing Multi Engine

CERTIFICATED
Type Certificated

AIR WORTHINESS DATE
2001-04-20

Engine Information

ENGINE TYPE
Turbo-FAN

NUMBER OF ENGINES
2

ENGINE MANUFACTURER
BMW ROLLS

ENGINE MODEL
BR 700 SERIES

Registration / Owner

REGISTRATION TYPE
Corporation

NAME
HAWAIIAN AIRLINES INC

ADDRESS
3375 KOAPAKA ST STE G350
HONOLULU, HI
968191804 US

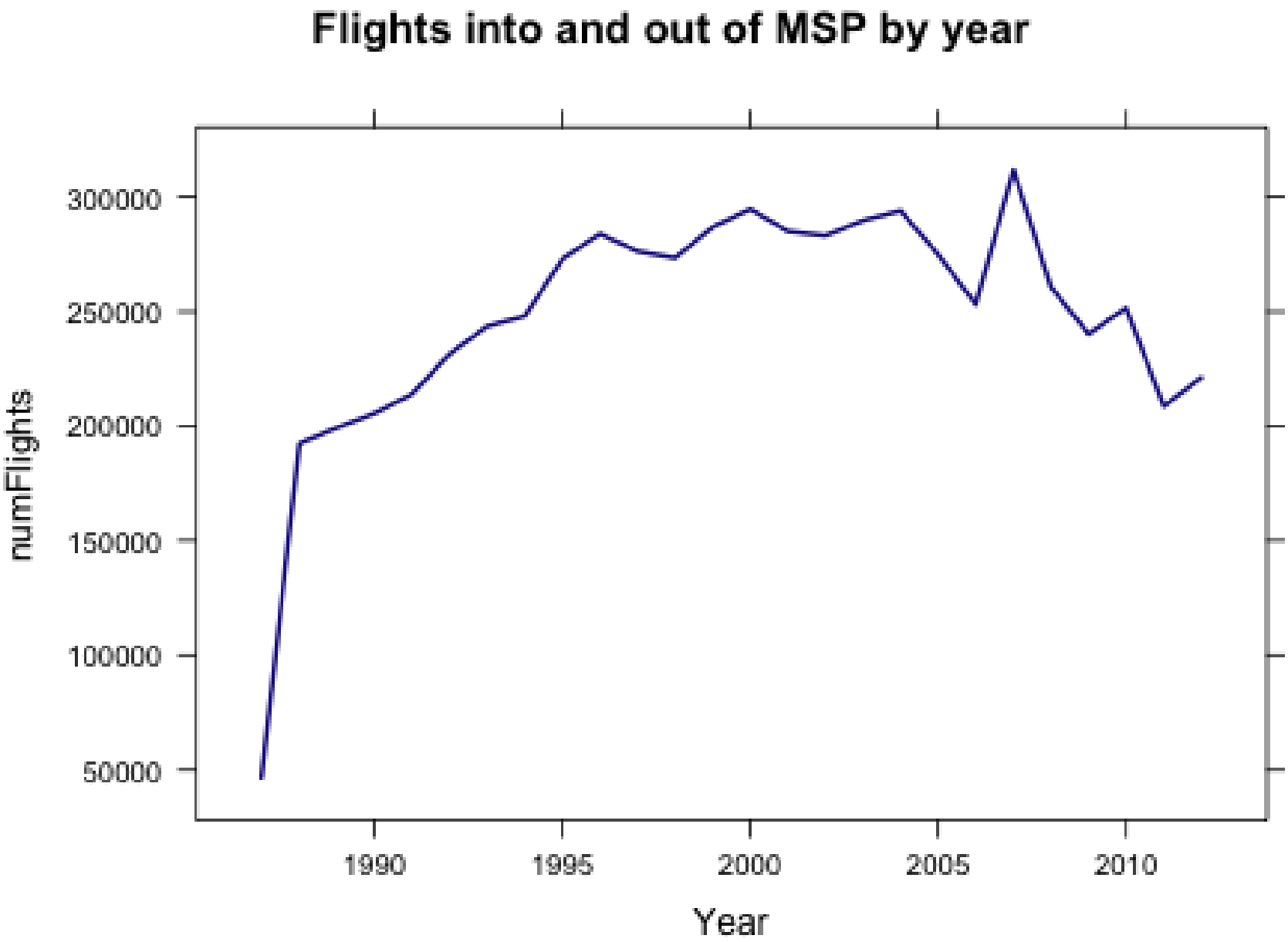
GROUP BY

```
dbGetQuery(con, "SELECT Year,
  COUNT(*) as numFlights FROM ontime GROUP BY Year")
   Year numFlights
1  1987    1311826
2  1988    5202096
3  1989    5041200
...
23 2009    6450285
24 2010    6450117
25 2011    6085281
26 2012    6096762
```

WHERE

```
dbGetQuery(con, "SELECT Year,
  COUNT(*) as numFlights FROM ontime
  WHERE (Dest='MSP' OR Origin='MSP') GROUP BY Year")
```

```
   Year numFlights
1  1987      46709
2  1988     192471
3  1989     199256
...
23 2009     240175
24 2010     251610
25 2011     208626
26 2012     221145
```


Flights into and out of MSP by year
numFlights
300000
250000
200000
150000
100000
50000
1990
1995
2000
2005
2010
Year

WHERE

```
dbGetQuery(con, "SELECT * FROM ontime
  WHERE (Origin='MSP' and Dest='BDL' AND Year=2012
    AND Month=10 AND DayofMonth=8)")
  Year Month DayofMonth DayOfWeek DepTime CRSDepTime ArrTime
1 2012    10          8         1     701        705    1038
2 2012    10          8         1    1319       1325    1659
3 2012    10          8         1    1922       1930    2255
  CRSArrTime UniqueCarrier FlightNum TailNum ActualElapsedTime
1       1043            EV      5545  N723EV                157
2       1659            DL      1226  N958DN                160
3       2305            DL      2170  N954DL                153
  CRSElapsedTime AirTime ArrDelay DepDelay Origin Dest Distance
1            158     134       -5       -4    MSP  BDL     1050
2            154     127        0       -6    MSP  BDL     1050
3            155     129      -10       -8    MSP  BDL     1050
  TaxiIn TaxiOut Cancelled CancellationCode Diverted
1      6      17         0                        0
2      6      27         0                        0
```

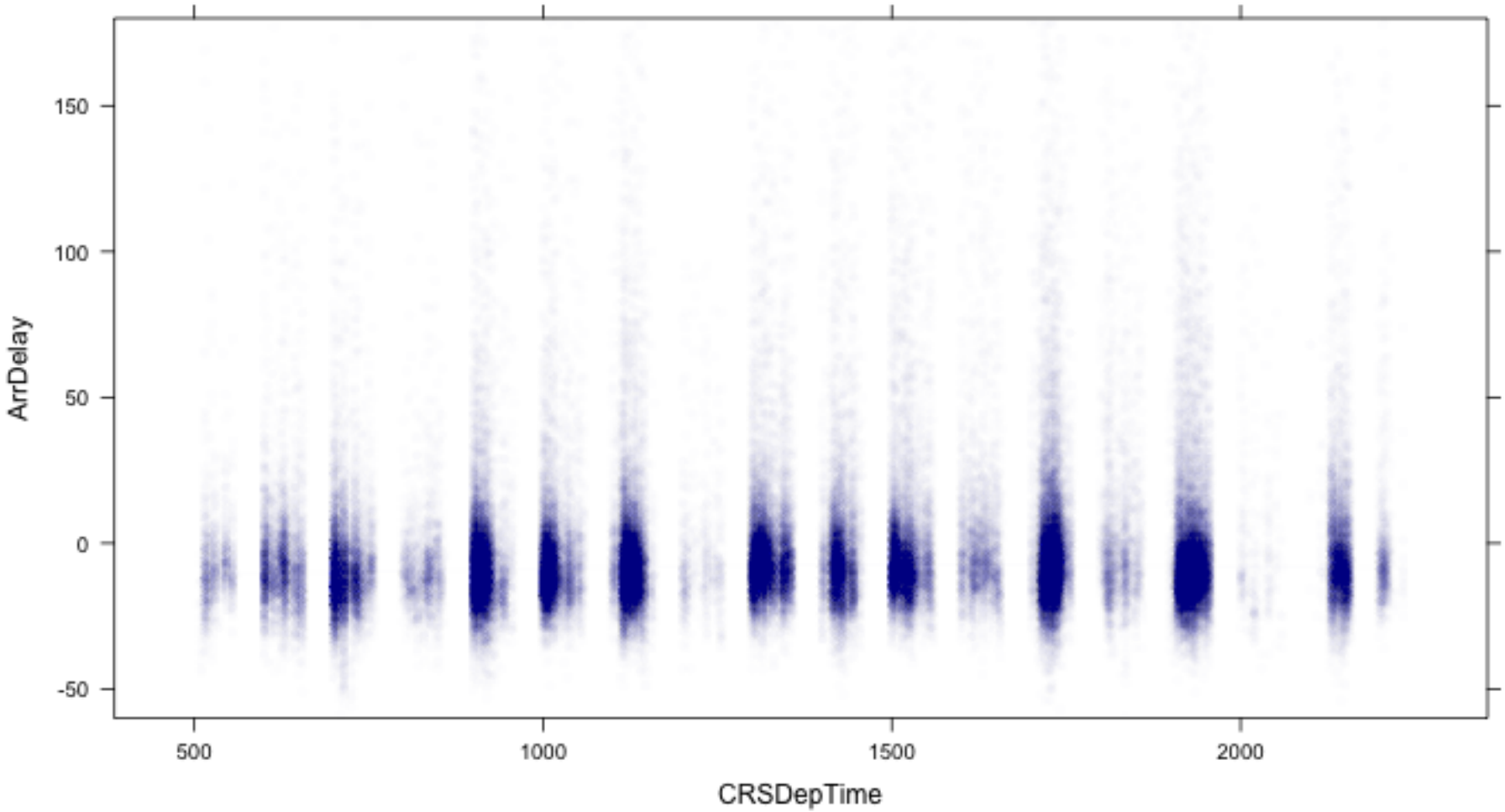
more complex selections

DL = Delta, EV = American Southeast

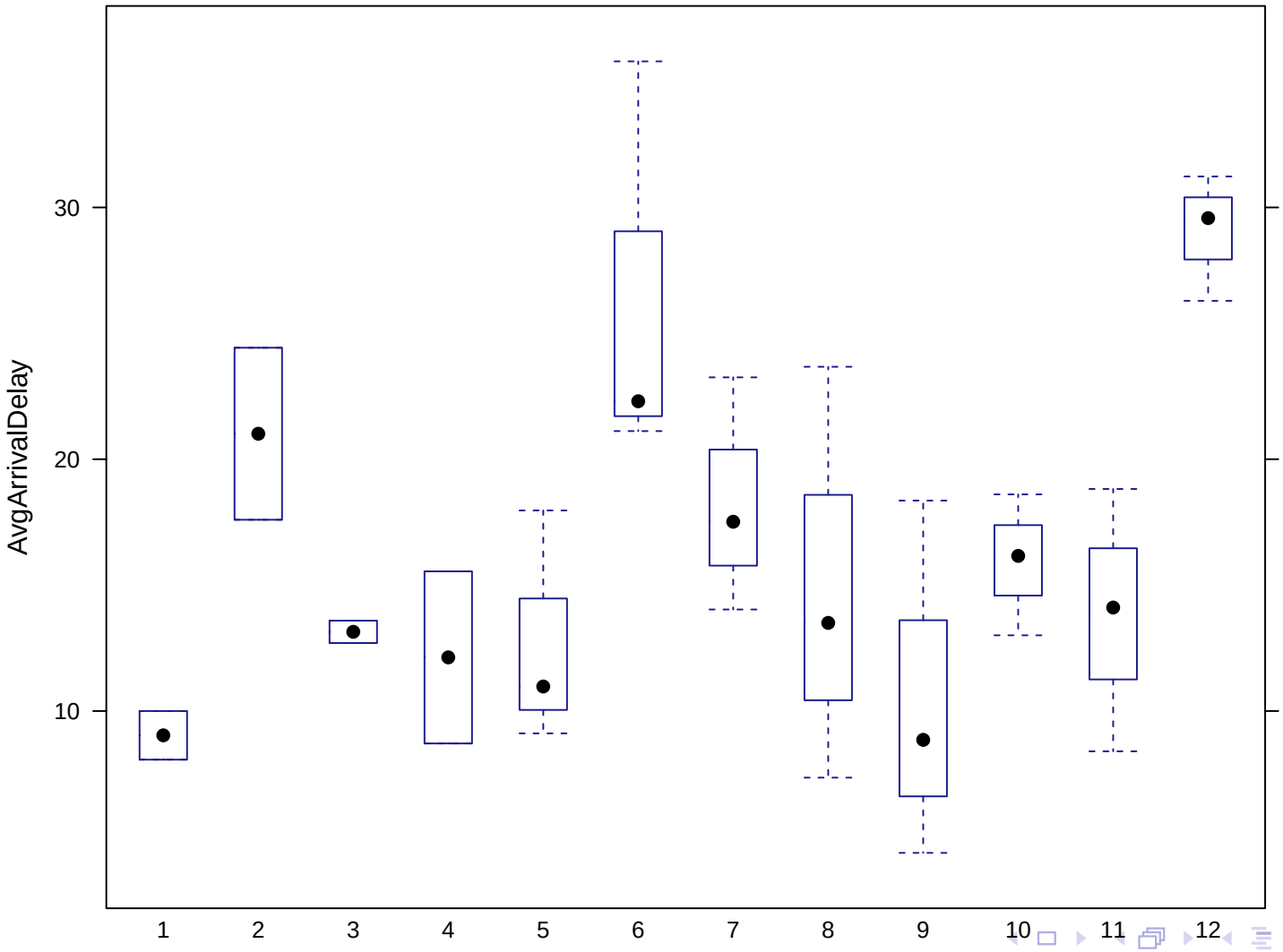
```
ds = dbGetQuery(con, "SELECT sum(1) as N, UniqueCarrier, Year,
  Month, DayofMonth, Dest, avg(if(ArrDelay< 0, 0, ArrDelay))
  as AvgArrivalDelay FROM ontime WHERE (Origin='MSP' and Dest='BDL'
    AND Year=2012 AND Month=10 AND DayofMonth=8)
  GROUP BY UniqueCarrier")
> ds
  N UniqueCarrier Year Month DayofMonth Dest AvgArrivalDelay
1 2            DL 2012    10          8  BDL               0
2 1            EV 2012    10          8  BDL               0
```

Arrival delay versus Departure Time

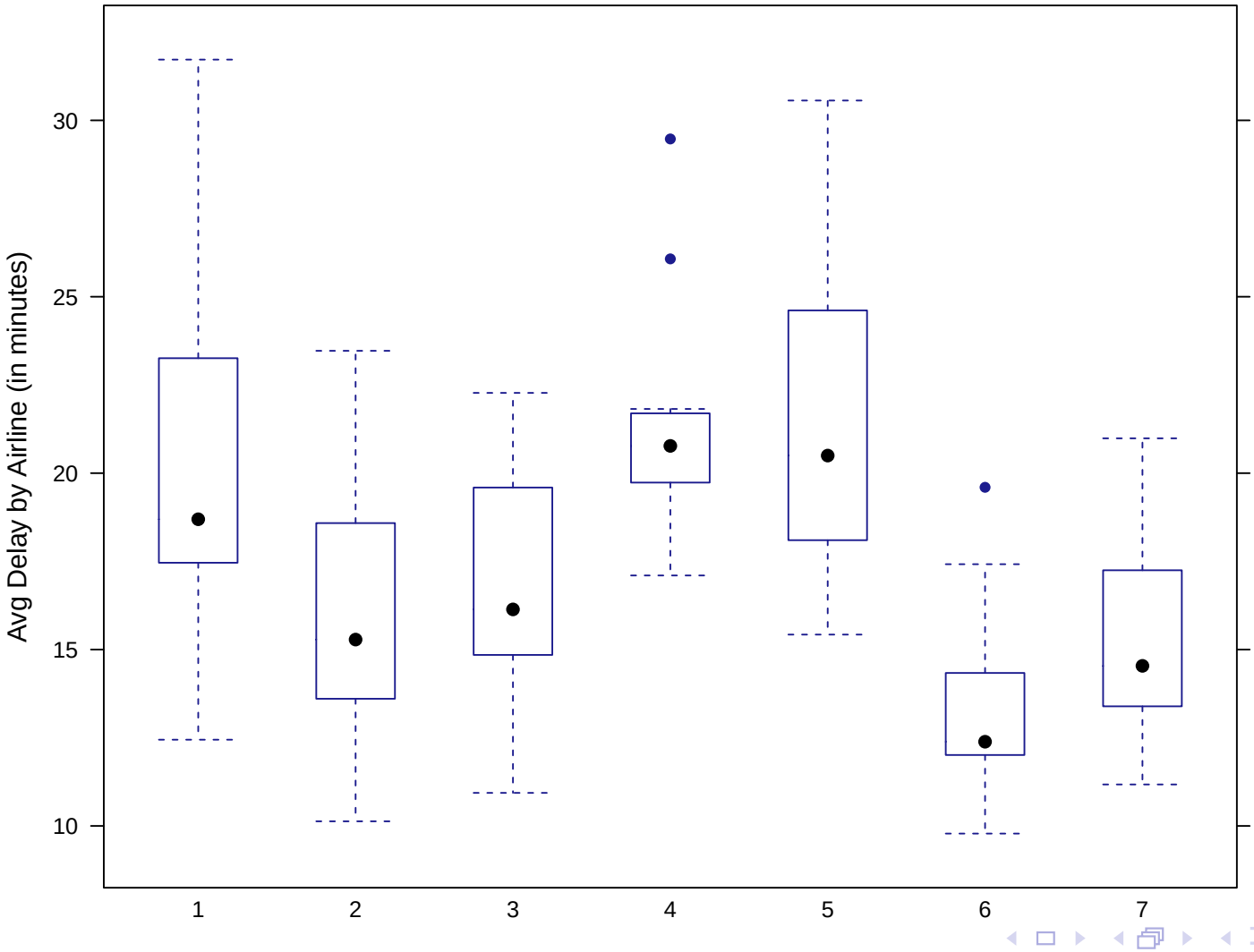
```
ds = dbGetQuery(con, "SELECT UniqueCarrier, Year, Month,
  DayofMonth, Origin, Dest, ArrDelay, CRSDepTime
  FROM ontime
  WHERE (Origin='MSP' AND Year=2012)")
> dim(ds)
[1] 110567      8
> xyplot(ArrDelay ~ CRSDepTime, alpha=0.01,
  ylim=c(-60,180), data=ds)
```


ArrDelay
150
100
50
0
-50
500
1000
1500
2000
CRSDepTime

Which month is it best to travel (airline averages/BDL)?



Which day is it best to travel (airline averages from BDL)?



Multiple tables

- so far all of our SELECTions have been from the single table `ontime`
- there's also a table for `carriers`

```
> dbGetQuery(con, "SELECT * FROM carriers LIMIT 6")
  code                            name
1  02Q                   Titan Airways
2  04Q             Tradewind Aviation
3  05Q           Comlux Aviation, AG
4  06Q Master Top Linhas Aereas Ltd.
5  07Q           Flair Airlines Ltd.
6  09Q                 Swift Air, LLC
```

Multiple tables

```
ds = dbGetQuery(con, "SELECT UniqueCarrier, FlightNum,
  Origin, Dest, DepTime, ArrTime FROM ontime WHERE
  Year = 1999 AND Month = 12 AND DayofMonth = 31 AND
  DepTime > ArrTime AND AirTime > 60;")
> head(ds)
  UniqueCarrier FlightNum Origin Dest DepTime ArrTime
1            DL        54    HNL  ATL    1759     706
2            DL       188    LAX  ATL    2230     515
3            DL       183    SLC  ATL    2045     203
4            DL      1946    SLC  ATL    2358     515
5            NW      1511    MSP  BIS    2308      27
6            UA       274    DEN  BOS    1837      12
```

Multiple tables

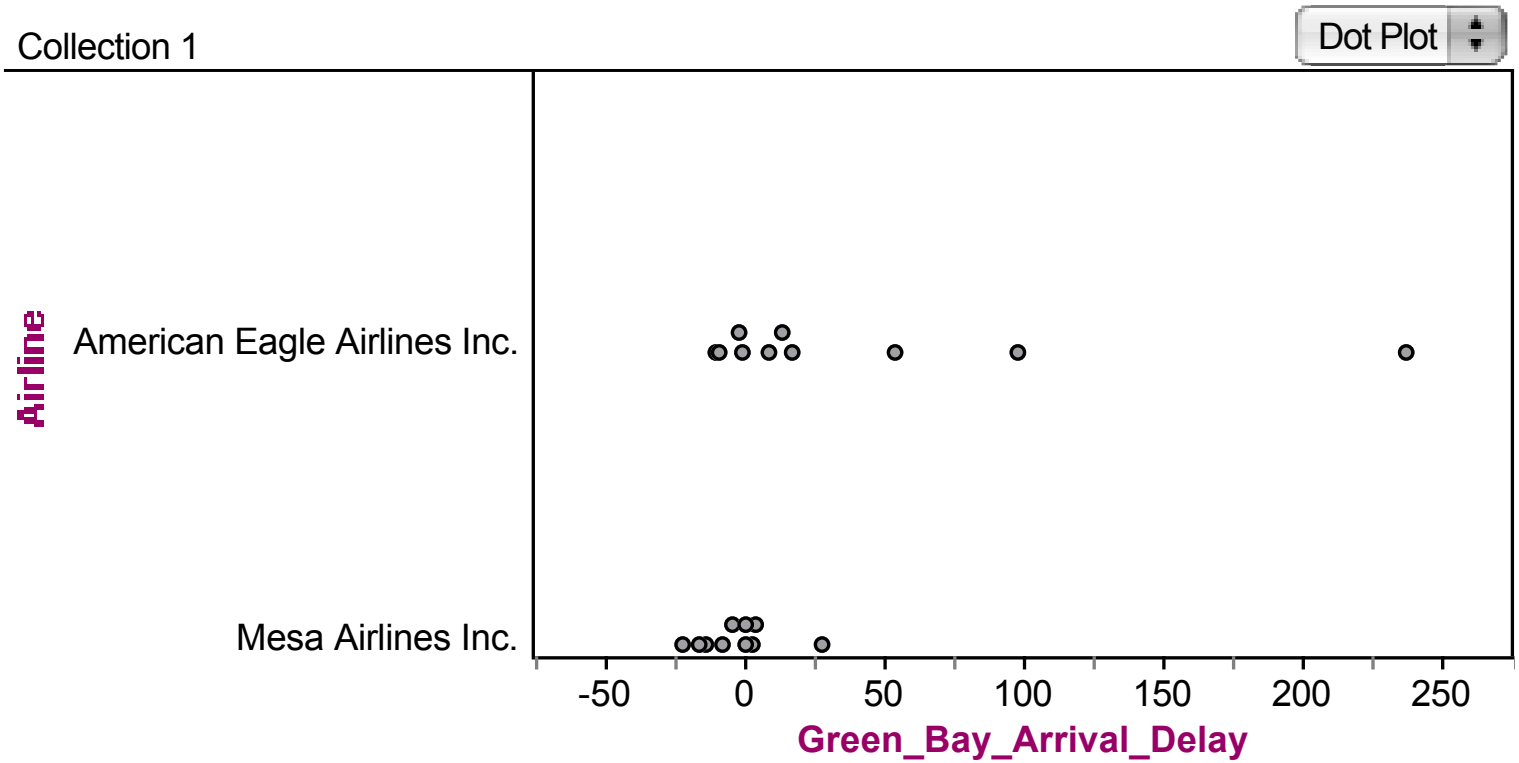
```
ds = dbGetQuery(con, "SELECT UniqueCarrier, c.name as
  CarrierName, FlightNum, Origin, Dest FROM ontime o
  LEFT JOIN carriers c ON o.UniqueCarrier = c.code
  WHERE Year = 1999 AND Month = 12 AND DayofMonth = 31
  AND DepTime > ArrTime AND AirTime > 60")
> head(ds)
UniqueCarrier                CarrierName FlightNum Origin Dest
1            DL    Delta Air Lines Inc.        54    HNL  ATL
2            DL    Delta Air Lines Inc.       188    LAX  ATL
3            DL    Delta Air Lines Inc.       183    SLC  ATL
4            DL    Delta Air Lines Inc.      1946    SLC  ATL
5            NW Northwest Airlines Inc.      1511    MSP  BIS
6            UA   United Air Lines Inc.       274    DEN  BOS
```

Using this in intro

start with model eliciting activity (how would you determine if one airline was more reliable than another?) using a small sample from one city pair (popularized by Garfield and Zieffler and colleagues)

1. Is there a difference in the reliability as measured by arrival time delays for these two regional airlines out of Chicago? Or are both airlines pretty much the same in terms of their arrival time delays?
2. If there are differences, are these differences consistent from city to city?
3. Are any differences you find large enough to influence travelers so that they are advised to choose one airline over the other (all other factors, like cost, being equal)?

Using this in intro



- have students determine when to "make a call"
- interpret differences in sample statistics between the airlines
- come up with a rule using two or more of those measures to determine when the "make a call"

Using this in intro

Data Values (in minutes)

AMERICAN EAGLE
-10 -9 -2 -1 9 13 17 54 98 236

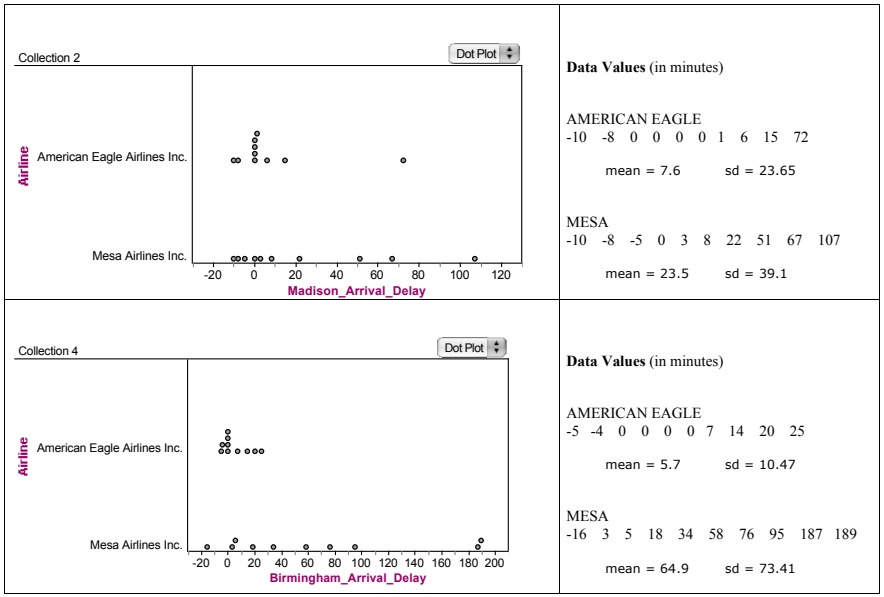
mean = 40.5 sd = 76.4

MESA
-22 -16 -14 -8 -5 0 0 3 4 28

mean = -3.0 sd = 13.92

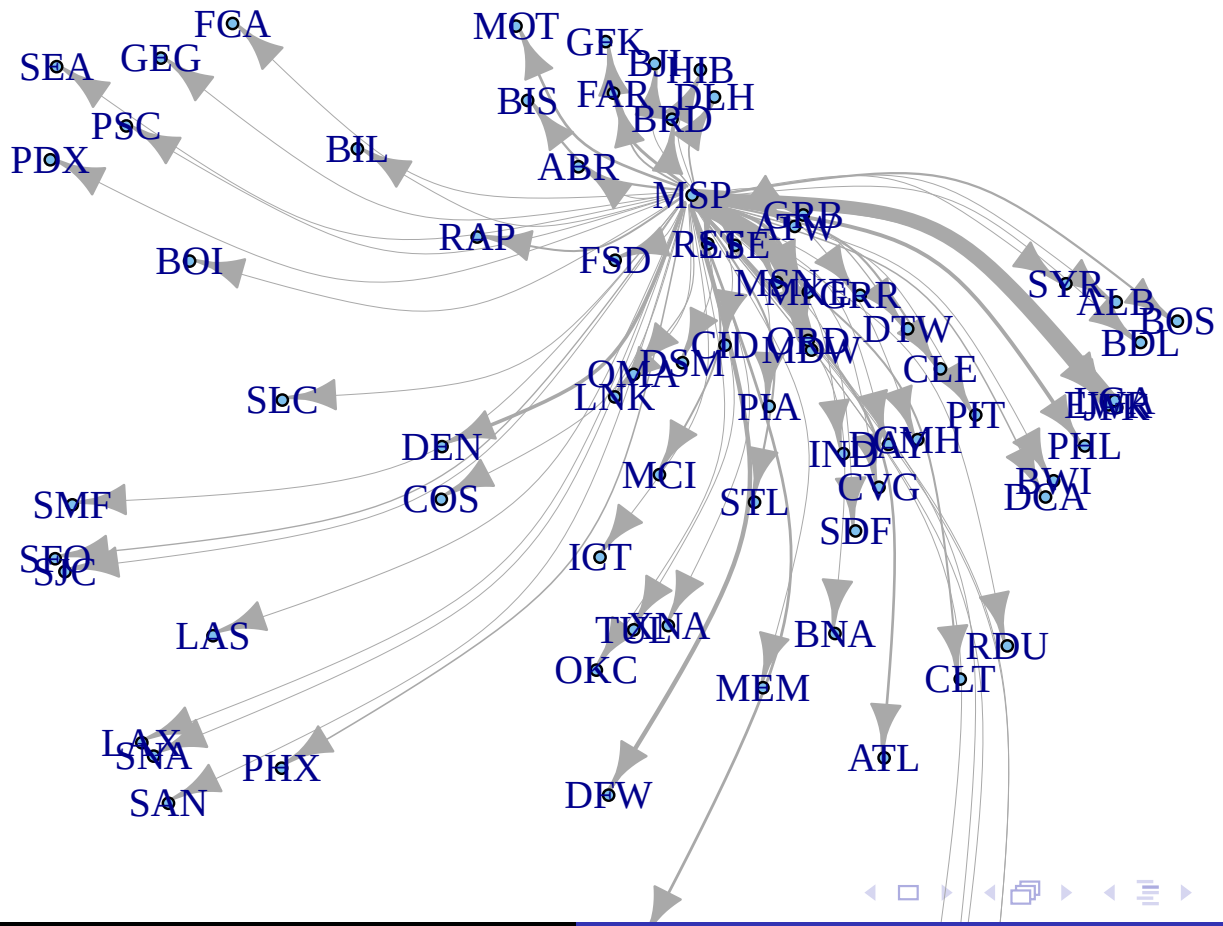
- have students determine when to "make a call"
- interpret differences in sample statistics between the airlines
- come up with a rule using two or more of those measures to determine when the "make a call"

Using this in intro



- compare to new city pairs (in class)
- return later in course to let them assess the performance of their rule (by repeatedly sampling)
- turn them loose to visualize and tell some new stories

Maps and visualization



Next steps

- more complex selections (see Lumley's work on doing this using R to simplify life for the user)
- multiple tables and more difficult joins
- more sophisticated tools in R (e.g. `plyr`)

Next steps

- need more knowledge of databases
- address issues of efficiency and performance
- big advantage of MySQL: caching of prior `SELECT` calls

> *The query cache stores the text of a SELECT statement together with the corresponding result that was sent to the client. If an identical statement is received later, the server retrieves the results from the query cache rather than parsing and executing the statement again. The query cache is shared among sessions, so a result set generated by one client can be sent in response to the same query issued by another client.*

Closing thoughts

- SQL is a powerful and flexible way to address big(ger) data
- straightforward to set up and use
- helps to bring more interesting data into the classroom

Activity

- download the file `PHX.Rmd` from `www.amherst.edu/~nhorton/stolaf`
- upload this to the RStudio server at `rstudio.stolaf.edu`
- run it and work through the questions listed there
- come up with your own questions and tell an interesting story

A beginner's guide to using SQL with R: database usage for fun and profit

Nicholas J. Horton

Amherst College, Amherst, MA, USA

October 8, 2013

nhorton@amherst.edu