Big data in the intro stats class: use of the airline delays dataset to expose students to a real-world, complex dataset

Nicholas J. Horton

Amherst College, Amherst, MA, USA

January 17, 2014

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`, examples at `http://www.amherst.edu/~nhorton/airlines`

Motivation and 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)

Motivation and 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)*

Motivation and 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 (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 (Gould, *Teaching Statistical Thinking in the Data Deluge*, 2014)
- 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

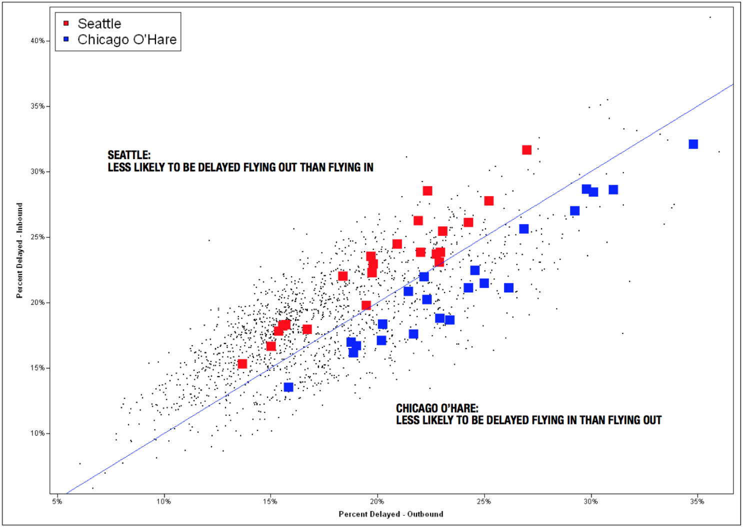
- dataset of flight arrival and departure details for all commercial flights within the USA, from October 1987 to April 2008 (but we now have through the end of 2012!)
- large dataset: nearly 150 million records
- aim: provide a graphical summary of important features of the data set
- winners presented 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

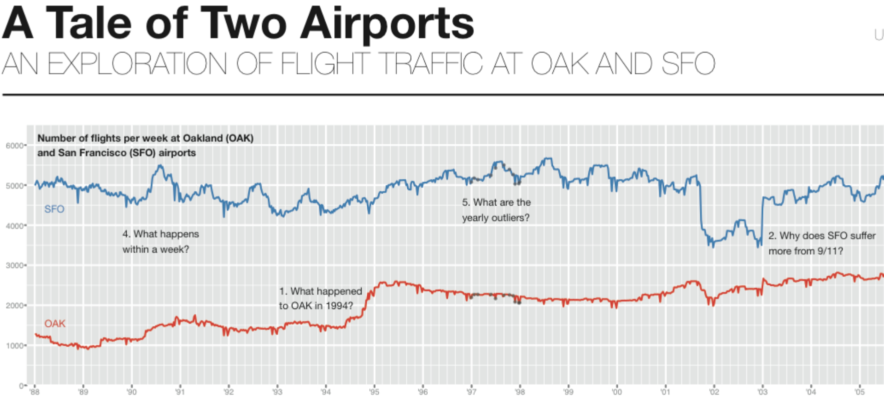
CAN WE SEE WHAT IS NOT THERE?
Planes have, for reasons such as maintenance, weather, or schedule fly empty between airports as so-called *Ghosts*. By tracking individual planes, we reveal their paths, including situations, where a plane lands in a different airport than where it takes off later, i.e. 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-
tance between airports of 1000 miles, corresponding to about 1.5 million

Sampling of the Data Expo 2009 winners



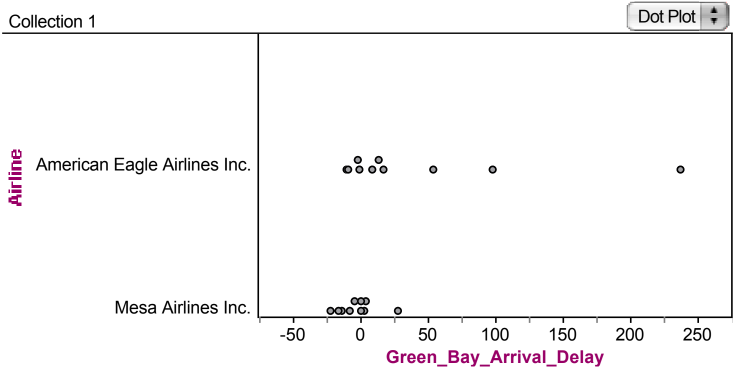
Model Eliciting Activity

- how would you determine if one airline was more reliable than another?
- give students a small sample from the airlines dataset for one city pair for two airlines
- CATALST project, `http://serc.carleton.edu/sp/library/mea/examples/example5.html`
- original MEA requires no technology
- students work in groups of 3 or 4

Statistical questions (when to "make the call")

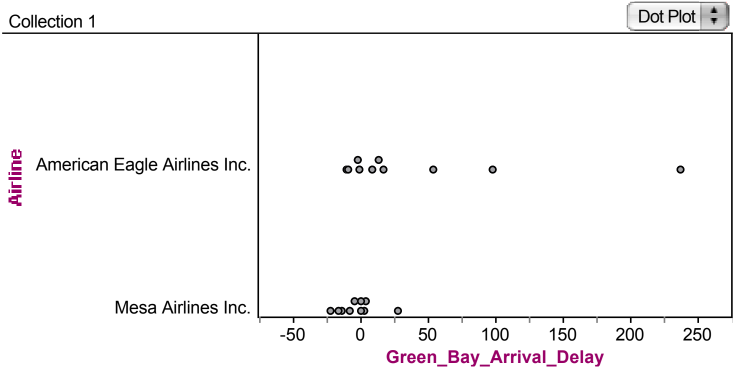
1. Is there a difference in the reliability as measured by arrival time delays for these two regional airlines out of Chicago? Or are both airlines pretty much the same in terms of their arrival time delays?
2. If there are differences, are these differences consistent from city to city?
3. Are any differences 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



- compare differences in 5 sample statistics
- come up with a rule using two or more of those measures to determine when the "make the call" for which airline might be more reliable

Using this in intro



- compare differences in 5 sample statistics
- examples: difference in mean delay, difference in proportion delayed, difference in IQR, difference in means for flights that were delayed

Using this in intro

Data Values (in minutes)

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

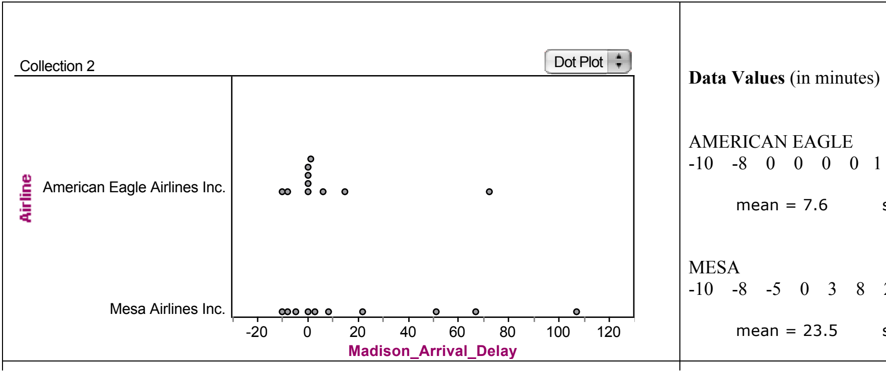
mean = 40.5 sd = 76.4

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

mean = -3.0 sd = 13.92

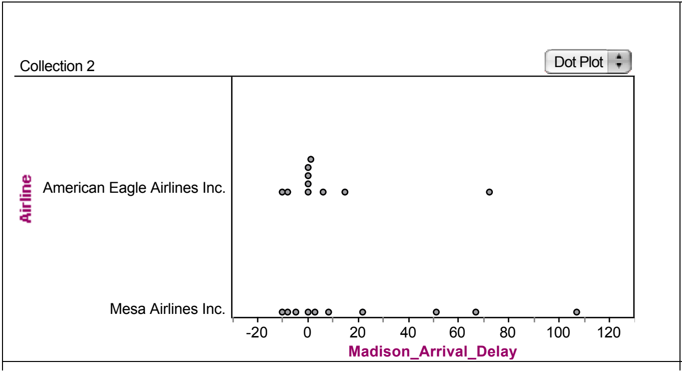
- compare differences in 5 sample statistics
- come up with a rule using two or more of those measures to determine when the "make the call" for which airline might be more reliable

Using this in intro



- compare to new city pairs (in class), and summarize performance of their rule

Using this in intro



Data Values (in minutes)

AMERICAN EAGLE
-10 -8 0 0 0 0 1

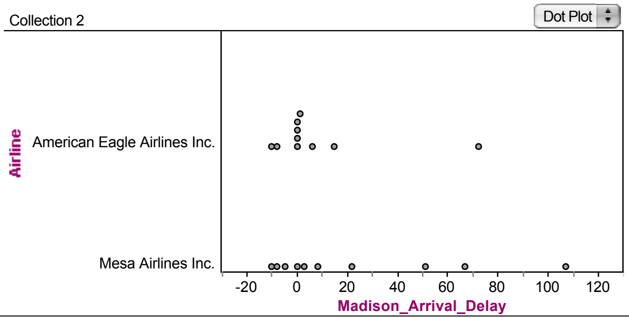
mean = 7.6

MESA
-10 -8 -5 0 3 8

mean = 23.5

- return later in course to let them assess the performance of their rule more formally (by repeatedly sampling)

Using this in intro



Data Values (in minutes)

AMERICAN EAGLE
-10 -8 0 0 0 0 1

mean = 7.6

MESA
-10 -8 -5 0 3 8

mean = 23.5

- explore further analyses and student generated questions (their favorate airline or airport) as part of end of semester project

Background on databases and SQL

- no technology needed for initial MEA
- modest investment can allow use of a rich dataset
- instructors need some 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 (can get up and running in an hour).
For a class, I'd recommend MySQL.

Creating the airline delays database (approx. 1 hour for SQLite)

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

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
```

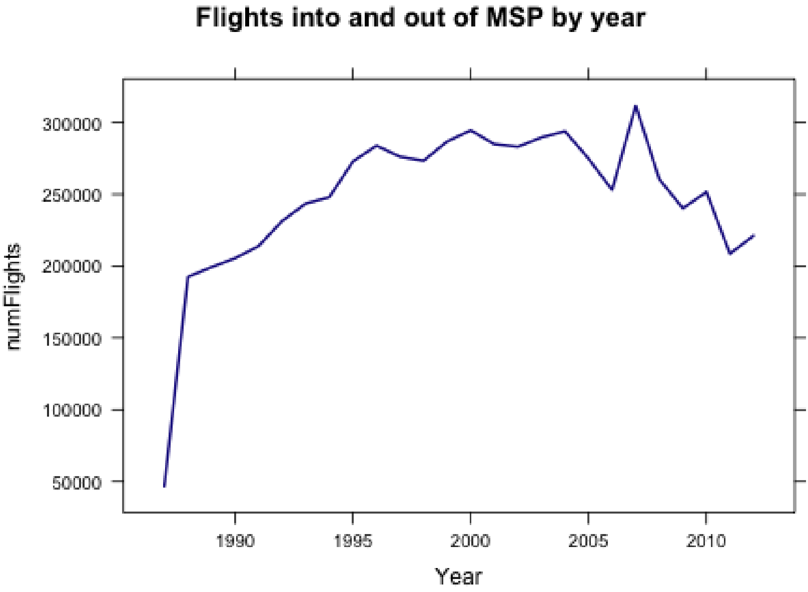
GROUP BY

```
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
```

Downloading data from Green Bay

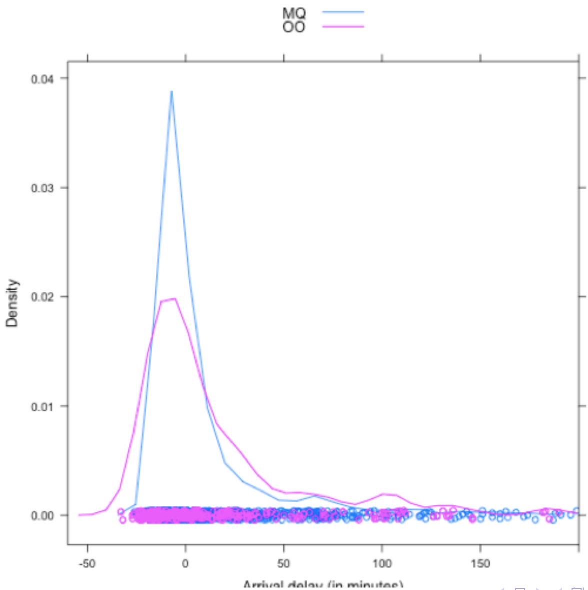
```
ds2 = dbGetQuery(con, "SELECT UniqueCarrier, ArrDelay,
   Month, Year, Origin, Dest FROM ontime
   WHERE Origin='GRB' AND Dest='ORD' AND Year=2005")
dim(ds2)
[1] 2166    6
```

Take a peek at 6 flights on Mesa

```
head(subset(ds2, UniqueCarrier == "MQ"))
##   UniqueCarrier ArrDelay Month Year Origin Dest
## 1            MQ       41     1 2005    GRB  ORD
## 2            MQ        3     1 2005    GRB  ORD
## 3            MQ      144     1 2005    GRB  ORD
## 4            MQ        9     1 2005    GRB  ORD
## 5            MQ      168     1 2005    GRB  ORD
## 6            MQ        5     1 2005    GRB  ORD
```

Testing using the population data

- once downloaded the distribution of flight delays by airline can be compared
- students can sample from this population, testing their rule each time
- requires a clear statement of their rule for the instructor or TA to craft a function in R
- example: one airline is better if it is at least 30 minutes less mean delays, and the sample standard deviation in each group is no more than 60 minutes


MQ
OO
Density
0.04
0.03
0.02
0.01
0.00
-50
0
50
100
150

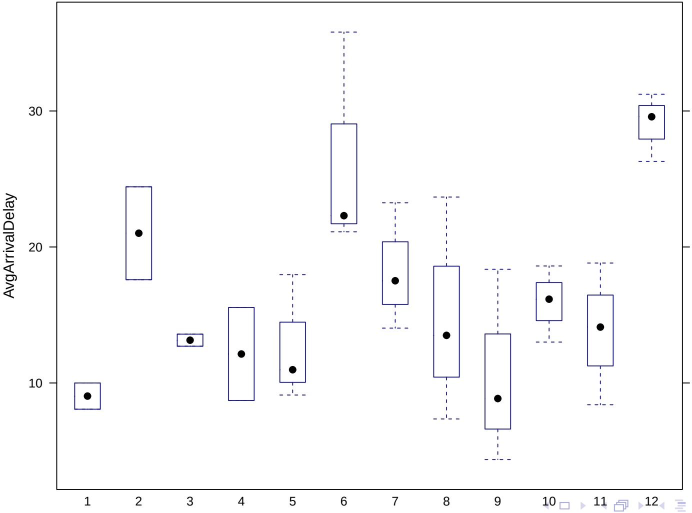
Comparison of rule over 2000 samples

```
res = do(2000) * compareI(sample(alleagle$ArrDelay, 10),
                          sample(allmesa$ArrDelay, 10))
tally(~result, data = res)
##
## Airline A Airline B   NEITHER     Total
##        12       183      1805      2000
```

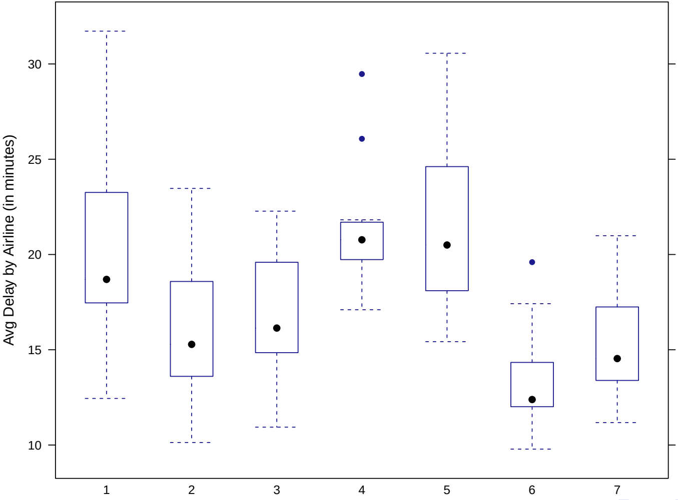
Closing thoughts

- MEA's bring big ideas into the classroom
- SQL is a powerful and flexible way to address big(ger) data
- straightforward to set up and use
- helps to allow instructors (and in later classes, students), tackle more interesting questions

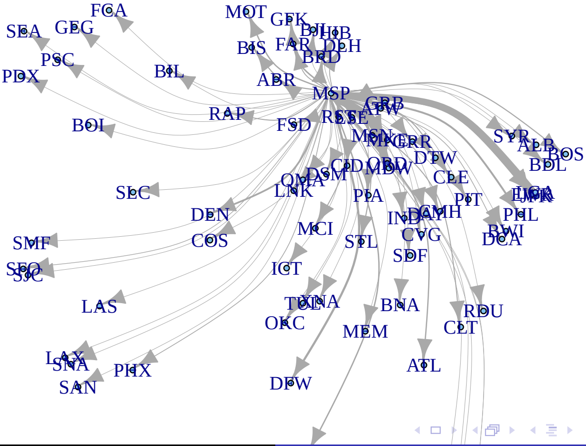
Which month is it best to travel (airline averages/BDL)?



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



Maps and visualization



Big data in the intro stats class: use of the airline delays dataset to expose students to a real-world, complex dataset

Nicholas J. Horton

Amherst College, Amherst, MA, USA

January 17, 2014

nhorton@amherst.edu

examples at `http://www.amherst.edu/~nhorton/airlines`