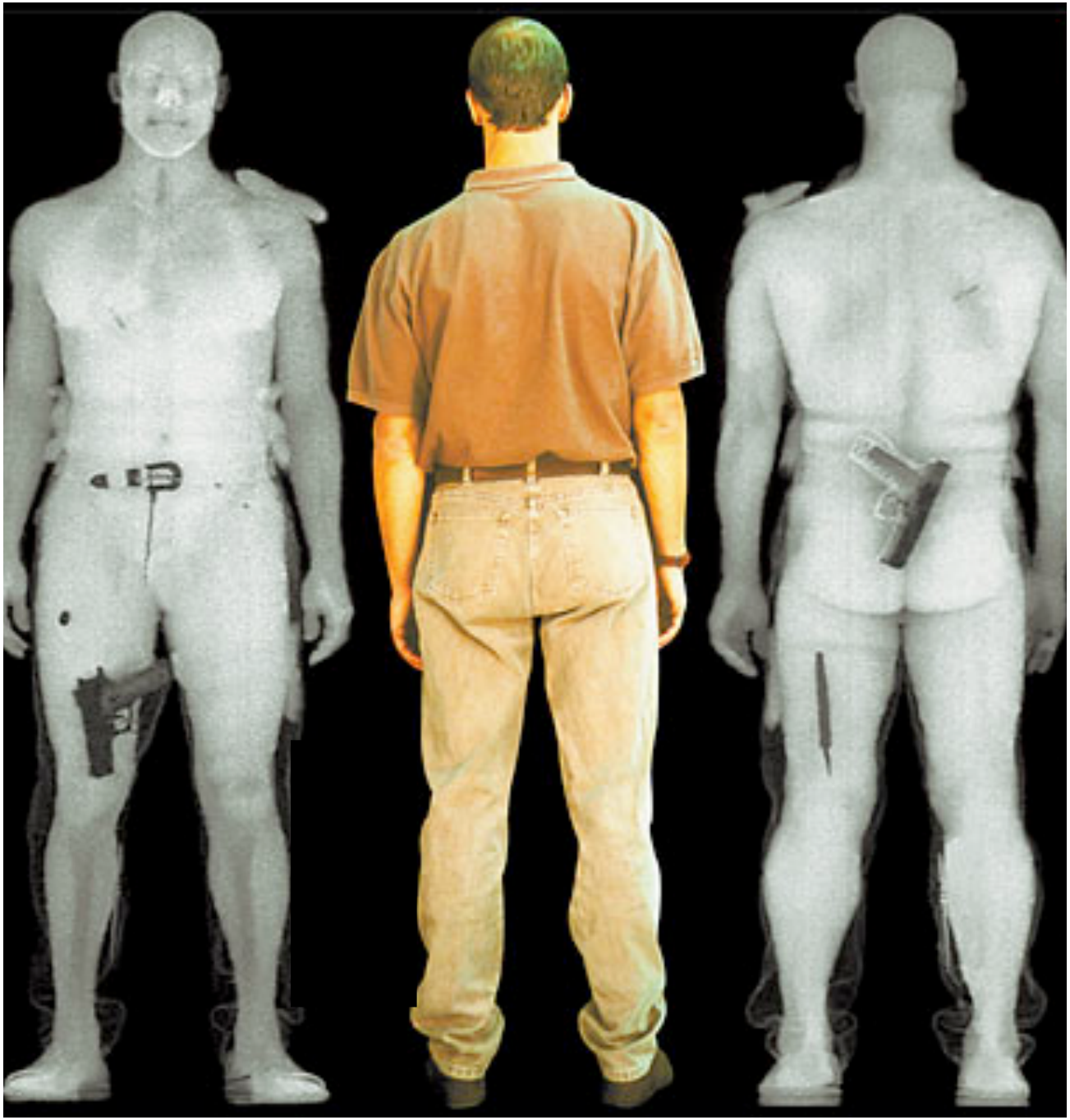


Last time

We played with the idea of a data table a fair bit -- We examined how we can reshape, aggregate and reformat data to provide us with alternate views of some phenomenon

Along the way, spent some time talking about privacy and about how the computer represents time -- We'll also learn some basic graphical tools for visualizing simple data types

Today

We start with another **“What could this amount to?”** example, this time using search engine logs as an input -- We again play privacy against utility of aggregate data

We then finish off the registrar’s data, examining a **“building view”** that starts us thinking about spatial data and anticipates where we’re headed later in the quarter -- We’ll also see some other data the registrar sold us and think about how **they could be usefully “joined” to address new questions**

Finally, we take up some **simple graphical and numerical summaries** (a la your Chapters 4 and 5) for a large survey data set from the CDC -- These data will be the subject of your next homework assignment (they also provide a modern “monitoring” or “scanning” example that builds on the Bills of Mortality)

What does it add up to?

Last time we saw that your geotagged Flickr images could be used to construct the outlines of places (cities, states, countries, continents) when looked at in the aggregate

We also mentioned a privacy issue related to search engines and that by considering all the searches you have performed over a long period of time, it might be possible to uniquely identify you

Today, we'll see what unexpected uses people can make of **your searching behavior** -- This is a medical or at least an epidemiological example!

Google Zeitgeist 2010

www.google.com/intl/en/press/zeitgeist2010/

Google zeitgeist

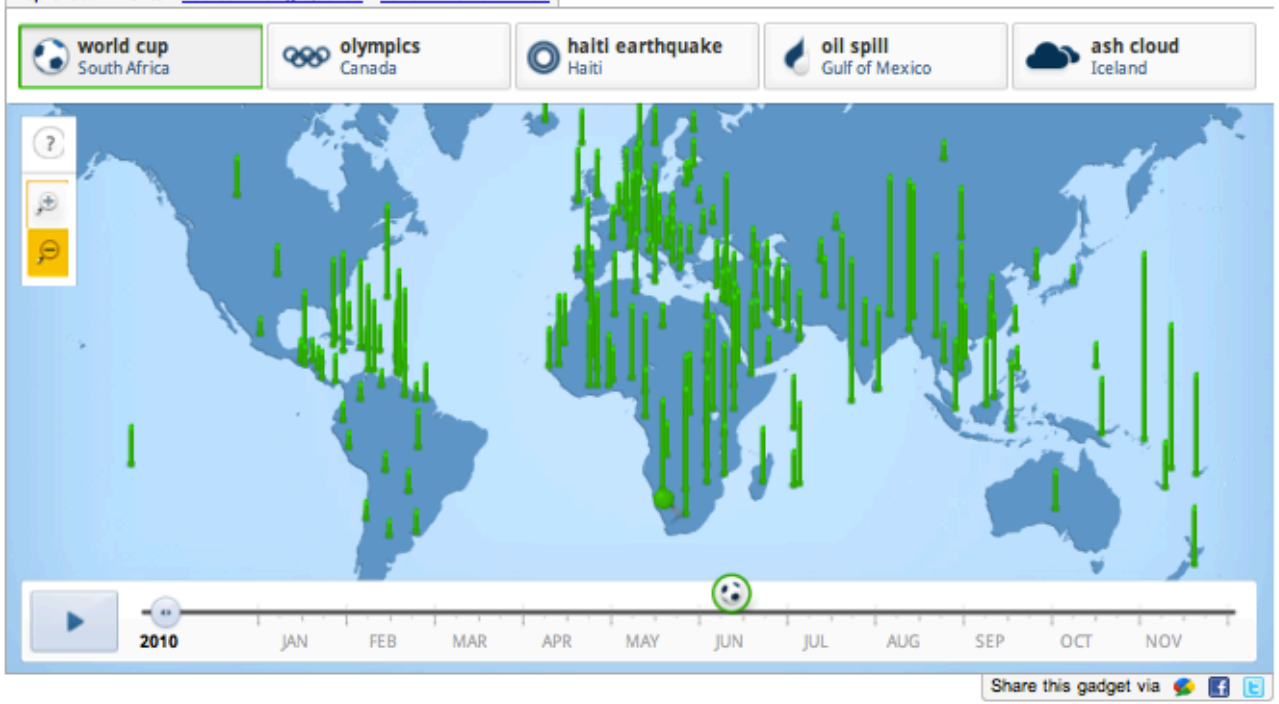
English

Global | United States | More Regions

Zeitgeist 2010: How the world searched

Based on the aggregation of billions of search queries people typed into Google this year, Zeitgeist captures the spirit of 2010.

Top Global Events · Fastest Rising Queries · 2010 in Review Video



Share this gadget via

Fastest Rising

1. chatroulette
2. ipad
3. justin bieber
4. nicki minaj
5. friv
6. myxer
7. katy perry
8. twitter
9. gamezer

Fastest Falling

1. swine flu
2. wamu
3. new moon
4. mininova
5. susan boyle
6. slumdog millionaire
7. circuit city
8. myspace layouts
9. michael jackson

Fastest Rising in Entertainment

1. justin bieber
2. shakira
3. eminem
4. netflix
5. youtube videos
6. lady gaga
7. kesha
8. nicki minaj
9. grooveshark

Flu season

Through the U.S. Influenza Sentinel Physicians Surveillance Network, the Centers for Disease Control and Prevention (CDC) monitors **the percentage of a physician's patients that exhibit Influenza-like Illness** (a fever and a cough and/or a sore throat, in the absence of a known cause other than the flu)

Physicians in the network send information to the CDC, which, in turn, aggregates the data across states and 10 or so higher-level regions (Pacific, Mountain, North East, etc.) -- Unfortunately this network can be **slow to aggregate report and slow to aggregate**

In terms of a surveillance system, Google (and Yahoo!) noticed that **people take to the web to self-diagnose when they're not feeling well**, and that this activity happens in advance of a trip to the doctor -- This means that **examining patterns in search queries might alert authorities to a coming epidemic** before it registers with the network of doctor

Google Trends: stuffy nose

www.google.com/trends?q=stuffy+nose&ctab=0&geo=all&date=all

Google trends

stuffy nose | Search Trends

Tip: Use commas to compare multiple search terms.

Sign in to see and export additional Trends data.

Searches | Websites

All regions | All years

stuffy nose

A Team of Egyptian doctors examines Yasser Arafat after he suffers nausea, fever, stuffy nose
San Diego Union Tribune - Oct 19 2004

B Condition Causes Chronic Stuffy Nose During Pregnancy
Milwaukee Channel.com - Jan 18 2005

C 'Stuffy Nose' Mouse: A Promise To Help Treat 31 Million With Sinusitis
Science Daily (press release) - Jul 23 2008

D Dealing with a stuffy nose
Calgary Herald - Oct 14 2010

E • The Stuffy Nose!
Sunday Standard - Nov 22 2010

F Stuffy Nose And Remedies For Quick Cure
Oneindia - Dec 23 2010

More news results »

Rank by: stuffy nose

Regions

1. United States
2. Canada
3. India
4. Australia
5. United Kingdom

Cities

1. Herndon, VA, USA
2. Waterloo, Canada
3. Ashburn, VA, USA
4. Reston, VA, USA
5. Pleasanton, CA, USA
6. Richardson, TX, USA
7. Boston, MA, USA
8. Toronto, Canada
9. New York, NY, USA
10. Los Angeles, CA, USA

Languages

1. English

Google Trends provides insights into broad search patterns. Please keep in mind that several approximations are used when computing these results.

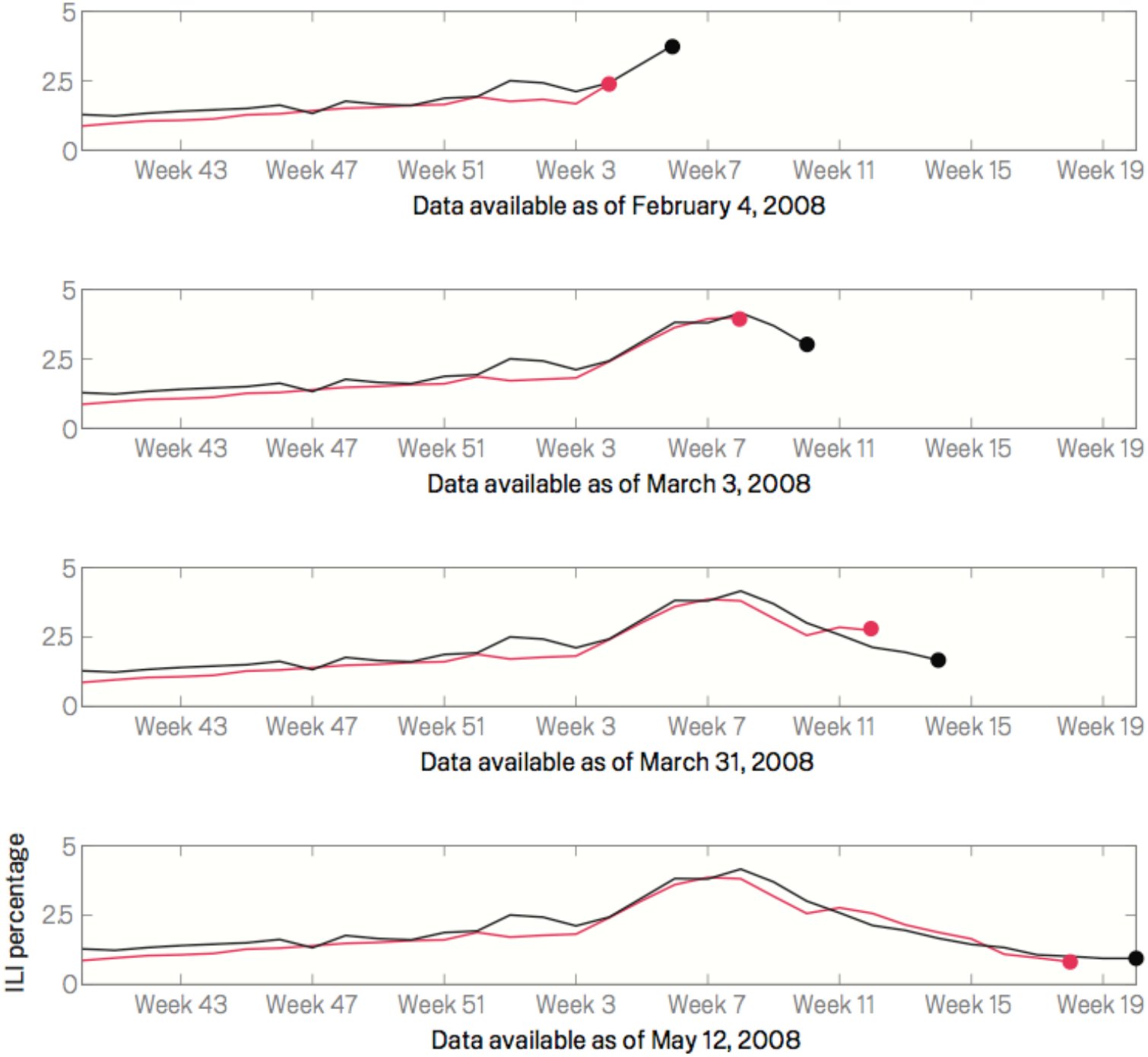
Google

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

[1]Google Inc. [2]Centers for Disease Control and Prevention

Epidemics of seasonal influenza are a major public health concern, causing tens of millions of respiratory illnesses and 250,000 to 500,000 deaths worldwide each year[1]. In addition to seasonal influenza, a new strain of influenza virus against which no prior immunity exists and that demonstrates human-to-human transmission could result in a pandemic with millions of fatalities[2]. Early detection of disease activity, when followed by a rapid response, can reduce the impact of both seasonal and pandemic influenza[3,4]. One way to improve early detection is to monitor health-seeking behavior in the form of online web search queries, which are submitted by millions of users around the world each day. Here we present a method of analyzing large numbers of Google search queries to track influenza-like illness in a population. Because the relative frequency of certain queries is highly correlated with the percentage of physician visits in which a patient presents with influenza-like symptoms, we can accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day. This approach may make it possible to utilize search queries to detect influenza epidemics in areas with a large population of web search users.


5
25
0
Week 43
Week 47
Week 51
Week 3
Week7
Week 11
Week 15
Week 19
Data available as of February 4, 2008
Week 7
Data available as of March 3, 2008
Data available as of March 31, 2008
ILI percentage
Data available as of May 12, 2008

Google Flu Trends

www.google.org/flutrends/

google.org Flu Trends

Language: English (United States)

Google.org home

Flu Trends

Home

Select country/regior

How does this work?

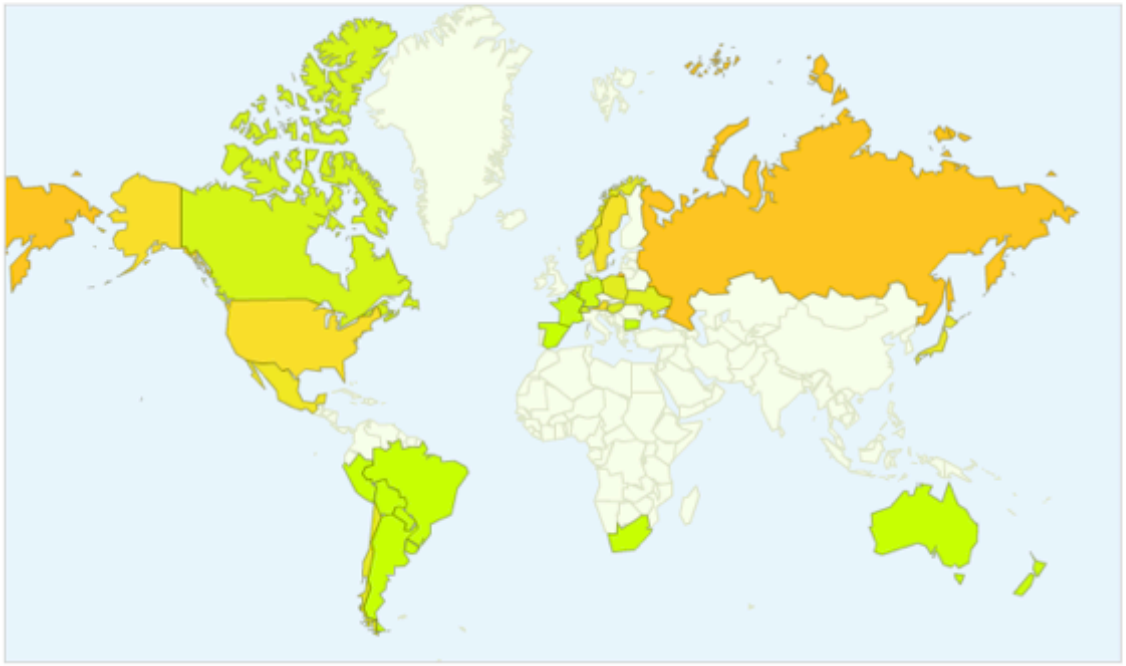
FAQ

Flu activity

- Intense
- High
- Moderate
- Low
- Minimal

Explore flu trends around the world

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »



Download world flu activity data - Animated flu trends for Google Earth - Compare flu trends across regions in Public Data Explorer

Google Flu Trends | United S

www.google.org/flutrends/us/#US

google.org Flu Trends

Language: English (United States)

Google.org home

Flu Trends

Home

United States

National

Download data

How does this work?

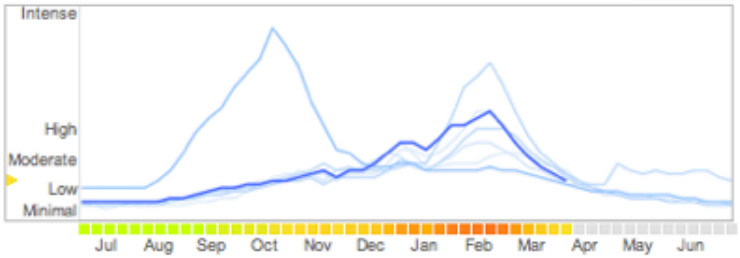
FAQ

Explore flu trends - United States

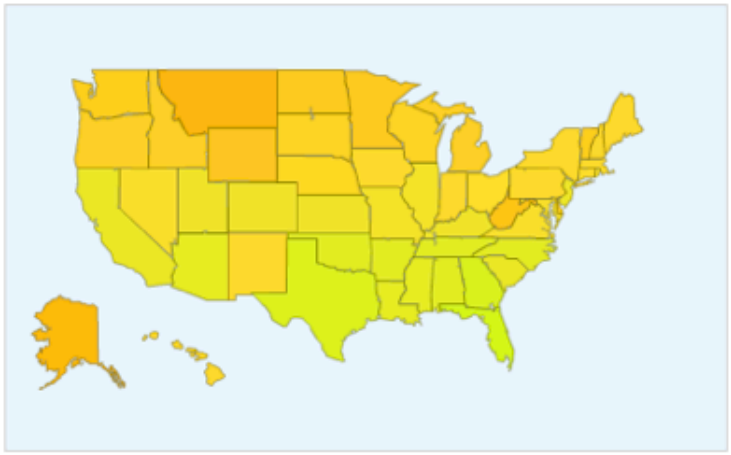
We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

National

2010-2011 Past years



States | Cities (Experimental)



Estimates were made using a model that proved accurate when compared to historic official flu activity data. Data current through April 2, 2011.

Fight influenza

CDC urges you to take these steps to protect yourself and others from the flu:

1. Get vaccinated against flu – it's your best defense.
2. Cover your cough, wash hands often.
3. Take antiviral drugs if your doctor recommends them.

CDC Centers for Disease Control and Prevention

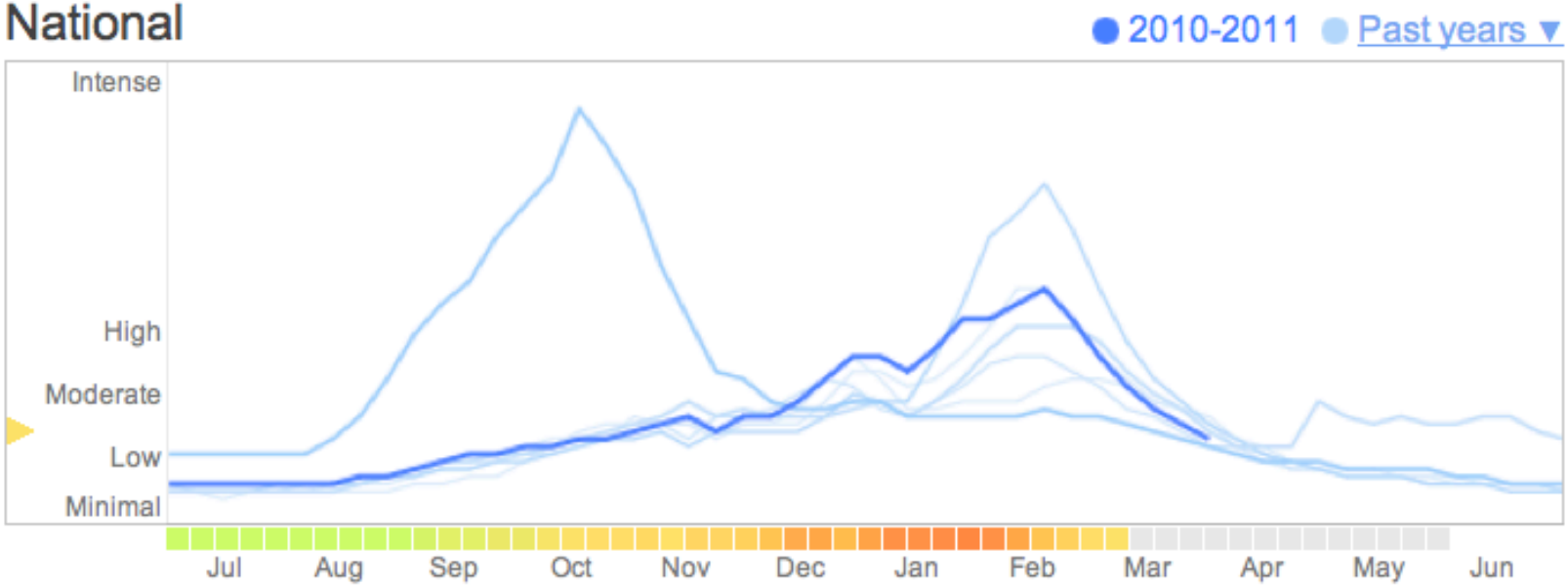
Animated Flu Trends in Google Earth

Download and explore Flu Trends data in Google Earth. Need Google Earth? Download it here.

Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

National



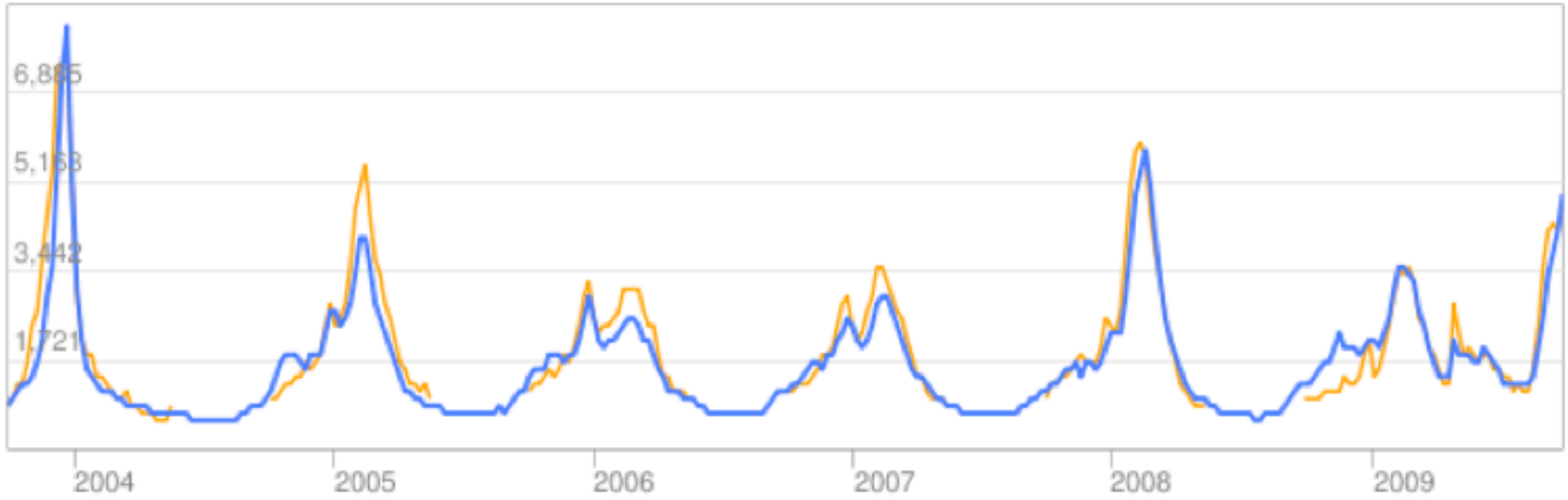
Historical estimates

See data for: United States

United States Flu Activity

Influenza estimate

Google Flu Trends estimate United States data



United States: Influenza-like illness (ILI) data provided publicly by the U.S. Centers for Disease Control.

Model selection

In all, Google reported testing out **50 million different search queries**, evaluating each with the simple regression equation (we'll get to this a little later in the quarter) -- Imagine a plot with the proportion of times a given term was queried on the x-axis and the CDC's ILI on the y-axis

In some sense, Google went through 50 million such plots (well, 450 million since they made one for each of 9 regions) and found those that had the "best" looking relationships -- That is**, those terms for which the search proportion predicted the ILI well**

Model validation

In the end, some 53 terms were chosen and a final model formed by averaging all the separate query ratios -- In an accompanying technical report the investigators write:

> *"We noted that the 53 highest scoring search queries appeared to be related to influenza-like illnesses. They describe symptoms, treatments, medications and other diseases that an average person might associate with influenza. The next highest scoring query 'high school basketball,' was the highest scoring off-topic query on the list: basketball season tends to coincide with influenza season in the United States."*

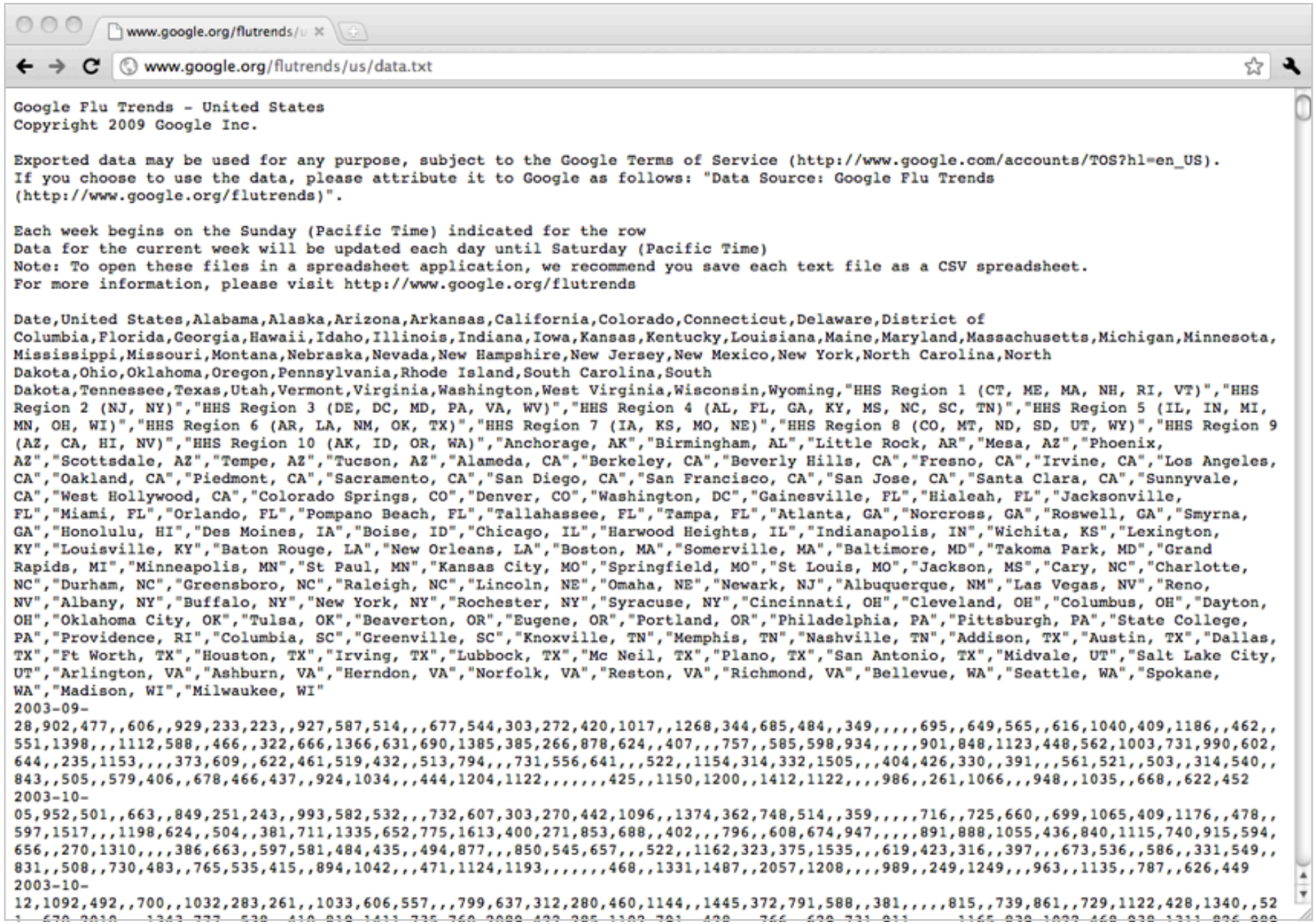
The model was then verified through the 2007-2008 flu season, with reportedly good success -- That is, the query model tracked the ILI reports, but were available much faster than the data from the physician's network

Flu season

Of course Google makes their estimates available (why would I be telling you this story otherwise?) -- The current data are in CSV (comma separated values) format and R has a “convenience” function for reading these data in

On the next page, we show the single line of code that reaches out to the web to pull these data -- notice that we are not using `source`() as we had in lab, but instead we are reading a CSV file from a URL

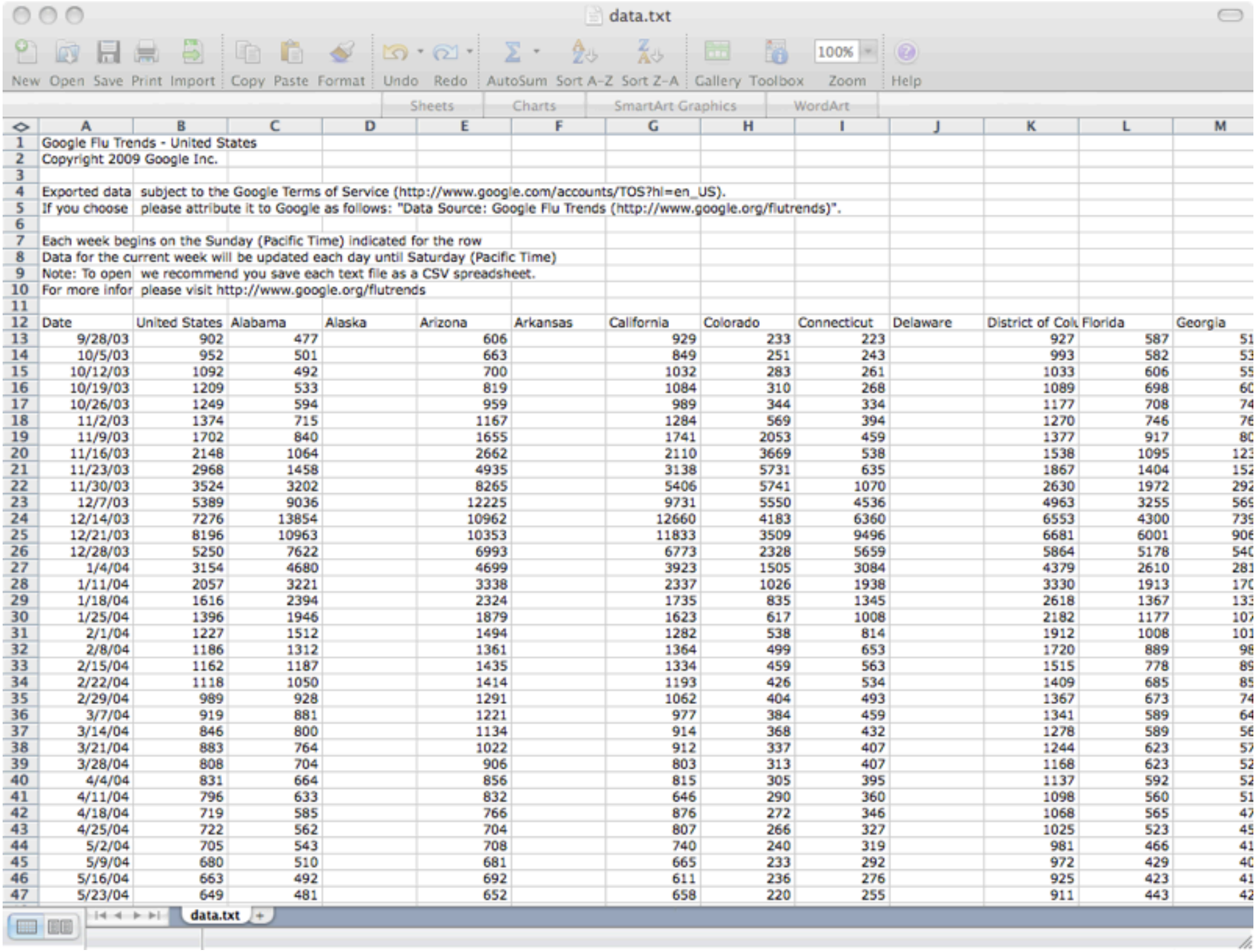
(Oh and the arguments to our read function tell R to skip the first 11 lines of the file, since there’s a lot of boilerplate up there; and that the file includes column headings)


www.google.org/flutrends/u
www.google.org/flutrends/us/data.txt
Google Flu Trends - United States
Copyright 2009 Google Inc.

Exported data may be used for any purpose, subject to the Google Terms of Service (http://www.google.com/accounts/TOS?hl=en_US).
If you choose to use the data, please attribute it to Google as follows: "Data Source: Google Flu Trends
(http://www.google.org/flutrends)".

Each week begins on the Sunday (Pacific Time) indicated for the row
Data for the current week will be updated each day until Saturday (Pacific Time)
Note: To open these files in a spreadsheet application, we recommend you save each text file as a CSV spreadsheet.
For more information, please visit http://www.google.org/flutrends

Date,United States,Alabama,Alaska,Arizona,Arkansas,California,Colorado,Connecticut,Delaware,District of
Columbia,Florida,Georgia,Hawaii,Idaho,Illinois,Indiana,Iowa,Kansas,Kentucky,Louisiana,Maine,Maryland,Massachusetts,Michigan,Minnesota,
Mississippi,Missouri,Montana,Nebraska,Nevada,New Hampshire,New Jersey,New Mexico,New York,North Carolina,North
Dakota,Ohio,Oklahoma,Oregon,Pennsylvania,Rhode Island,South Carolina,South
Dakota,Tennessee,Texas,Utah,Vermont,Virginia,Washington,West Virginia,Wisconsin,Wyoming,"HHS Region 1 (CT, ME, MA, NH, RI, VT)","HHS
Region 2 (NJ, NY)","HHS Region 3 (DE, DC, MD, PA, VA, WV)","HHS Region 4 (AL, FL, GA, KY, MS, NC, SC, TN)","HHS Region 5 (IL, IN, MI,
MN, OH, WI)","HHS Region 6 (AR, LA, NM, OK, TX)","HHS Region 7 (IA, KS, MO, NE)","HHS Region 8 (CO, MT, ND, SD, UT, WY)","HHS Region 9
(AZ, CA, HI, NV)","HHS Region 10 (AK, ID, OR, WA)","Anchorage, AK","Birmingham, AL","Little Rock, AR","Mesa, AZ","Phoenix,
AZ","Scottsdale, AZ","Tempe, AZ","Tucson, AZ","Alameda, CA","Berkeley, CA","Beverly Hills, CA","Fresno, CA","Irvine, CA","Los Angeles,
CA","Oakland, CA","Piedmont, CA","Sacramento, CA","San Diego, CA","San Francisco, CA","San Jose, CA","Santa Clara, CA","Sunnyvale,
CA","West Hollywood, CA","Colorado Springs, CO","Denver, CO","Washington, DC","Gainesville, FL","Hialeah, FL","Jacksonville,
FL","Miami, FL","Orlando, FL","Pompano Beach, FL","Tallahassee, FL","Tampa, FL","Atlanta, GA","Norcross, GA","Roswell, GA","Smyrna,
GA","Honolulu, HI","Des Moines, IA","Boise, ID","Chicago, IL","Harwood Heights, IL","Indianapolis, IN","Wichita, KS","Lexington,
KY","Louisville, KY","Baton Rouge, LA","New Orleans, LA","Boston, MA","Somerville, MA","Baltimore, MD","Takoma Park, MD","Grand
Rapids, MI","Minneapolis, MN","St Paul, MN","Kansas City, MO","Springfield, MO","St Louis, MO","Jackson, MS","Cary, NC","Charlotte,
NC","Durham, NC","Greensboro, NC","Raleigh, NC","Lincoln, NE","Omaha, NE","Newark, NJ","Albuquerque, NM","Las Vegas, NV","Reno,
NV","Albany, NY","Buffalo, NY","New York, NY","Rochester, NY","Syracuse, NY","Cincinnati, OH","Cleveland, OH","Columbus, OH","Dayton,
OH","Oklahoma City, OK","Tulsa, OK","Beaverton, OR","Eugene, OR","Portland, OR","Philadelphia, PA","Pittsburgh, PA","State College,
PA","Providence, RI","Columbia, SC","Greenville, SC","Knoxville, TN","Memphis, TN","Nashville, TN","Addison, TX","Austin, TX","Dallas,
TX","Ft Worth, TX","Houston, TX","Irving, TX","Lubbock, TX","Mc Neil, TX","Plano, TX","San Antonio, TX","Midvale, UT","Salt Lake City,
UT","Arlington, VA","Ashburn, VA","Herndon, VA","Norfolk, VA","Reston, VA","Richmond, VA","Bellevue, WA","Seattle, WA","Spokane,
WA","Madison, WI","Milwaukee, WI"
2003-09-
28,902,477,,606,,929,233,223,,927,587,514,,,677,544,303,272,420,1017,,1268,344,685,484,,349,,,,,695,,649,565,,616,1040,409,1186,,462,,
551,1398,,,1112,588,,466,,322,666,1366,631,690,1385,385,266,878,624,,407,,,757,,585,598,934,,,,,901,848,1123,448,562,1003,731,990,602,
644,,235,1153,,,,373,609,,622,461,519,432,,513,794,,,731,556,641,,,522,,1154,314,332,1505,,,404,426,330,,391,,,561,521,,503,,314,540,,
843,,505,,579,406,,678,466,437,,924,1034,,,444,1204,1122,,,,,,,425,,1150,1200,,1412,1122,,,,986,,261,1066,,,948,,1035,,668,,622,452
2003-10-
05,952,501,,663,,849,251,243,,993,582,532,,,732,607,303,270,442,1096,,1374,362,748,514,,359,,,,,716,,725,660,,699,1065,409,1176,,478,,
597,1517,,,1198,624,,504,,381,711,1335,652,775,1613,400,271,853,688,,402,,,796,,608,674,947,,,,,891,888,1055,436,840,1115,740,915,594,
656,,270,1310,,,,386,663,,597,581,484,435,,494,877,,,850,545,657,,,522,,1162,323,375,1535,,,619,423,316,,397,,,673,536,,586,,331,549,,
831,,508,,730,483,,765,535,415,,894,1042,,,471,1124,1193,,,,,,,468,,1331,1487,,2057,1208,,,,989,,249,1249,,,963,,1135,,787,,626,449
2003-10-
12,1092,492,,700,,1032,283,261,,1033,606,557,,,799,637,312,280,460,1144,,1445,372,791,588,,381,,,,,815,,739,861,,729,1122,428,1340,,52



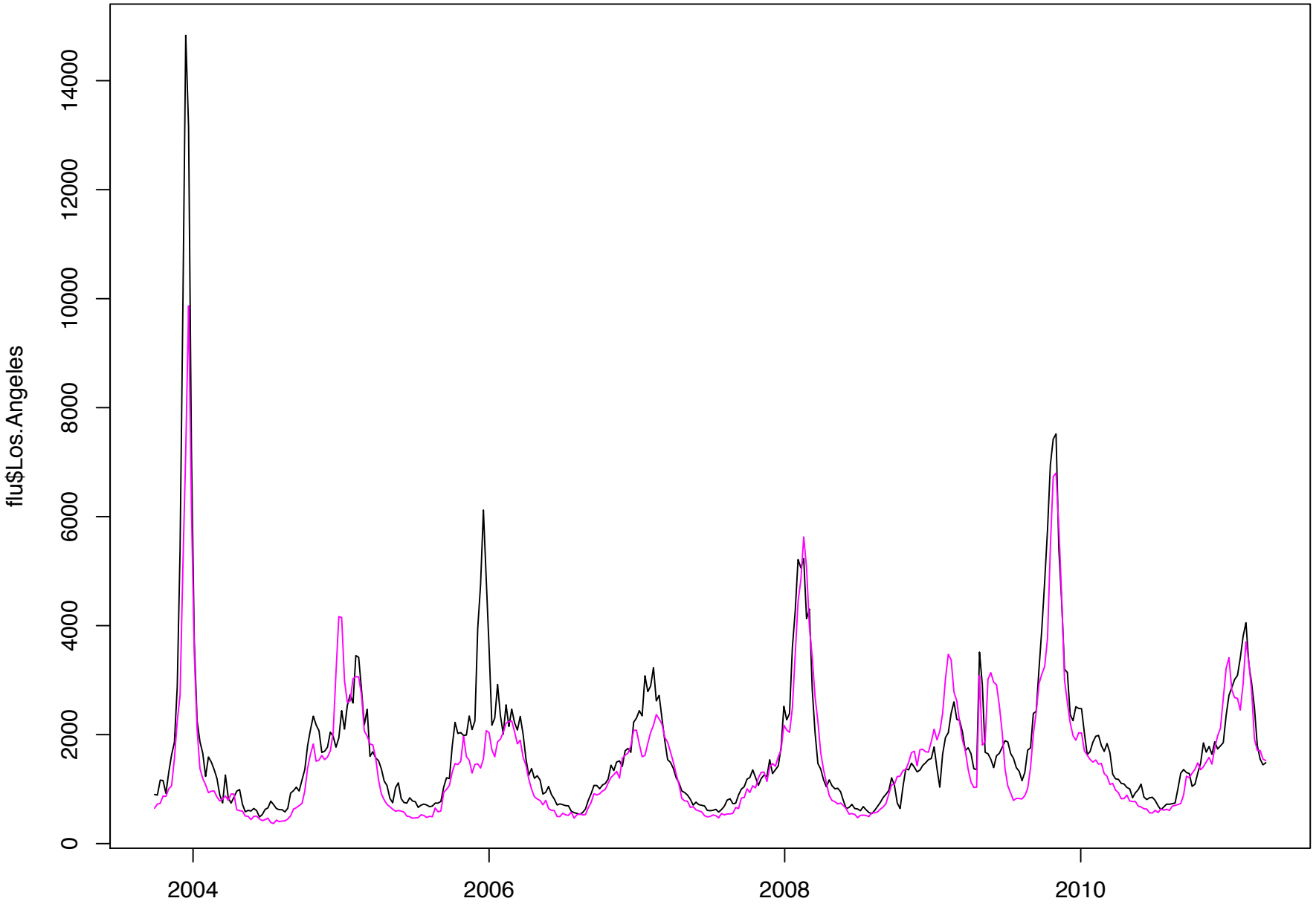
Google Flu Trends - United States
Copyright 2009 Google Inc.

Exported data subject to the Google Terms of Service (http://www.google.com/accounts/TOS?hl=en_US).
If you choose please attribute it to Google as follows: "Data Source: Google Flu Trends (http://www.google.org/flutrends)".

Each week begins on the Sunday (Pacific Time) indicated for the row
Data for the current week will be updated each day until Saturday (Pacific Time)
Note: To open we recommend you save each text file as a CSV spreadsheet.
For more infor please visit http://www.google.org/flutrends

Date	United States	Alabama	Alaska	Arizona	Arkansas	California	Colorado	Connecticut	Delaware	District of Col	Florida	Georgia
9/28/03	902	477		606		929	233	223		927	587	51
10/5/03	952	501		663		849	251	243		993	582	53
10/12/03	1092	492		700		1032	283	261		1033	606	55
10/19/03	1209	533		819		1084	310	268		1089	698	60
10/26/03	1249	594		959		989	344	334		1177	708	74
11/2/03	1374	715		1167		1284	569	394		1270	746	76
11/9/03	1702	840		1655		1741	2053	459		1377	917	80
11/16/03	2148	1064		2662		2110	3669	538		1538	1095	123
11/23/03	2968	1458		4935		3138	5731	635		1867	1404	152
11/30/03	3524	3202		8265		5406	5741	1070		2630	1972	292
12/7/03	5389	9036		12225		9731	5550	4536		4963	3255	569
12/14/03	7276	13854		10962		12660	4183	6360		6553	4300	739
12/21/03	8196	10963		10353		11833	3509	9496		6681	6001	906
12/28/03	5250	7622		6993		6773	2328	5659		5864	5178	540
1/4/04	3154	4680		4699		3923	1505	3084		4379	2610	281
1/11/04	2057	3221		3338		2337	1026	1938		3330	1913	170
1/18/04	1616	2394		2324		1735	835	1345		2618	1367	133
1/25/04	1396	1946		1879		1623	617	1008		2182	1177	107
2/1/04	1227	1512		1494		1282	538	814		1912	1008	101
2/8/04	1186	1312		1361		1364	499	653		1720	889	98
2/15/04	1162	1187		1435		1334	459	563		1515	778	89
2/22/04	1118	1050		1414		1193	426	534		1409	685	85
2/29/04	989	928		1291		1062	404	493		1367	673	74
3/7/04	919	881		1221		977	384	459		1341	589	64
3/14/04	846	800		1134		914	368	432		1278	589	56
3/21/04	883	764		1022		912	337	407		1244	623	57
3/28/04	808	704		906		803	313	407		1168	623	52
4/4/04	831	664		856		815	305	395		1137	592	52
4/11/04	796	633		832		646	290	360		1098	560	51
4/18/04	719	585		766		876	272	346		1068	565	47
4/25/04	722	562		704		807	266	327		1025	523	45
5/2/04	705	543		708		740	240	319		981	466	41
5/9/04	680	510		681		665	233	292		972	429	40
5/16/04	663	492		692		611	236	276		925	423	41
5/23/04	649	481		652		658	220	255		911	443	42

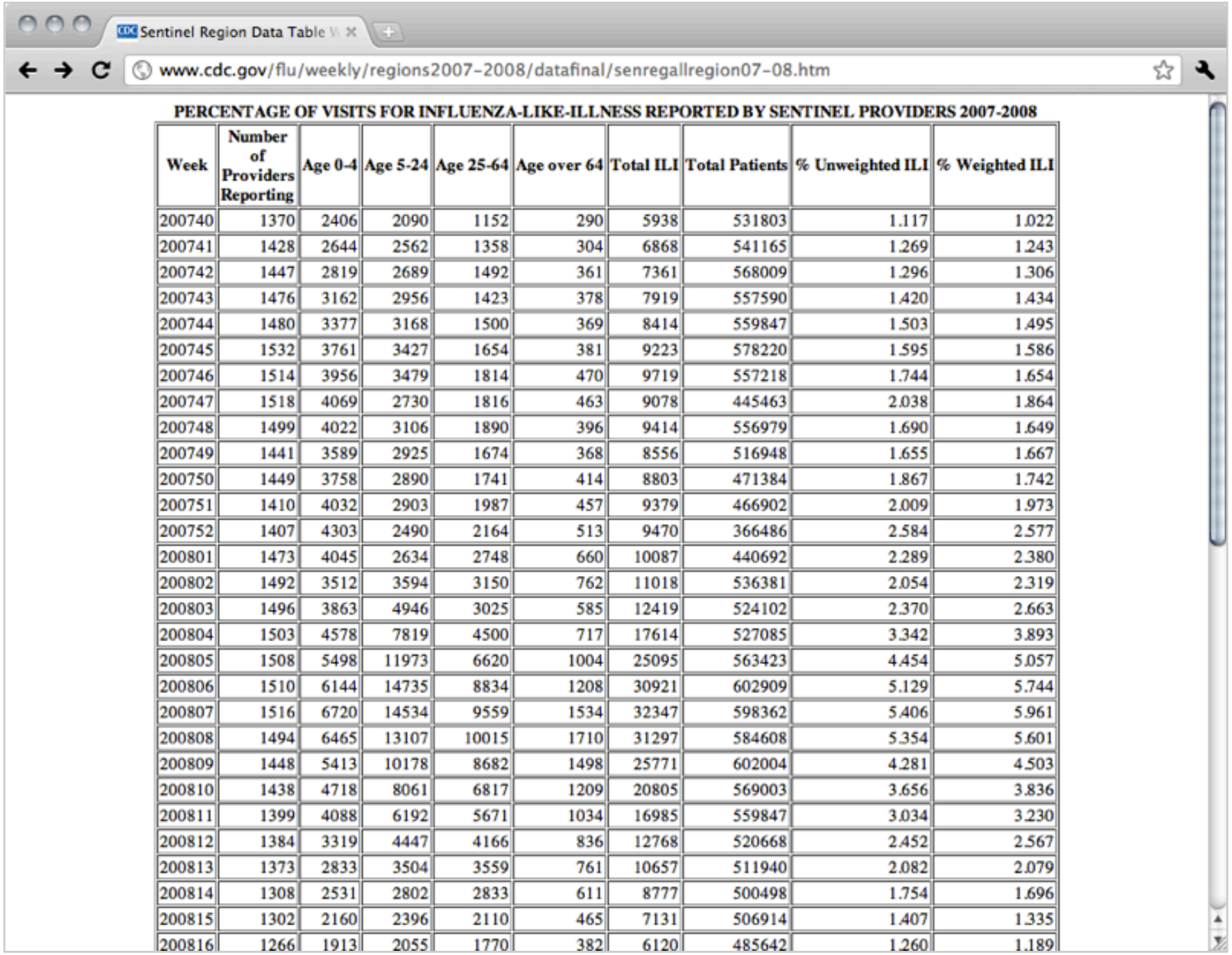
```
> flu <- read.csv(url("http://www.google.org/flutrends/us/data.txt"),skip=11)
> dim(flu)
[1] 393 180

> names(flu)
  [1] "Date"
  [2] "United.States"
  [3] "Alabama"
  [4] "Alaska"
  [5] "Arizona"
  [6] "Arkansas"
  [7] "California"
  [8] "Colorado"
  ...                    (clipped output)
 [71] "Tucson..AZ"
 [72] "Alameda..CA"
 [73] "Berkeley..CA"
 [74] "Beverly.Hills..CA"
 [75] "Fresno..CA"
 [76] "Irvine..CA"
 [77] "Los.Angeles..CA"
  ...                    (clipped output)
[135] "Albany..NY"
[136] "Buffalo..NY"
[137] "New.York..NY"
[138] "Rochester..NY"
[139] "Syracuse..NY"
 ...                    (clipped output)

> plot(as.Date(flu$Date),flu$Los.Angeles,type="l")
> lines(as.Date(flu$Date),flu$New.York,col="magenta")
```


flu$Los.Angeles
0
2000
4000
6000
8000
10000
12000
14000
2004
2006
2008
2010

One final note

The CDC also makes its ILI data available -- On the next page we present their data for the 2007-2008 flu season (it was hard to find newer data on their web site although I'm reasonably sure it has been published)

The data are in the form of an HTML table and require other R-related tools to read it in -- Even when we read it in, **do you see some issues we might face when comparing it to the Google data release?**



PERCENTAGE OF VISITS FOR INFLUENZA-LIKE-ILLNESS REPORTED BY SENTINEL PROVIDERS 2007-2008

Week	Number of Providers Reporting	Age 0-4	Age 5-24	Age 25-64	Age over 64	Total ILI	Total Patients	% Unweighted ILI	% Weighted ILI
200740	1370	2406	2090	1152	290	5938	531803	1.117	1.022
200741	1428	2644	2562	1358	304	6868	541165	1.269	1.243
200742	1447	2819	2689	1492	361	7361	568009	1.296	1.306
200743	1476	3162	2956	1423	378	7919	557590	1.420	1.434
200744	1480	3377	3168	1500	369	8414	559847	1.503	1.495
200745	1532	3761	3427	1654	381	9223	578220	1.595	1.586
200746	1514	3956	3479	1814	470	9719	557218	1.744	1.654
200747	1518	4069	2730	1816	463	9078	445463	2.038	1.864
200748	1499	4022	3106	1890	396	9414	556979	1.690	1.649
200749	1441	3589	2925	1674	368	8556	516948	1.655	1.667
200750	1449	3758	2890	1741	414	8803	471384	1.867	1.742
200751	1410	4032	2903	1987	457	9379	466902	2.009	1.973
200752	1407	4303	2490	2164	513	9470	366486	2.584	2.577
200801	1473	4045	2634	2748	660	10087	440692	2.289	2.380
200802	1492	3512	3594	3150	762	11018	536381	2.054	2.319
200803	1496	3863	4946	3025	585	12419	524102	2.370	2.663
200804	1503	4578	7819	4500	717	17614	527085	3.342	3.893
200805	1508	5498	11973	6620	1004	25095	563423	4.454	5.057
200806	1510	6144	14735	8834	1208	30921	602909	5.129	5.744
200807	1516	6720	14534	9559	1534	32347	598362	5.406	5.961
200808	1494	6465	13107	10015	1710	31297	584608	5.354	5.601
200809	1448	5413	10178	8682	1498	25771	602004	4.281	4.503
200810	1438	4718	8061	6817	1209	20805	569003	3.656	3.836
200811	1399	4088	6192	5671	1034	16985	559847	3.034	3.230
200812	1384	3319	4447	4166	836	12768	520668	2.452	2.567
200813	1373	2833	3504	3559	761	10657	511940	2.082	2.079
200814	1308	2531	2802	2833	611	8777	500498	1.754	1.696
200815	1302	2160	2396	2110	465	7131	506914	1.407	1.335
200816	1266	1913	2055	1770	382	6120	485642	1.260	1.189

The registrar's data (fin)

Last time, we examined alternate views of the registrar's data, taking us from a view focused on "**enrollment events**" (a student in a class) to a view in which **students** were the units of observation

In thinking over the uses for these data, many of you suggested another view, one that focuses on **the use of the buildings on campus** -- Here we have an opportunity to consider building usage both in time as well as space

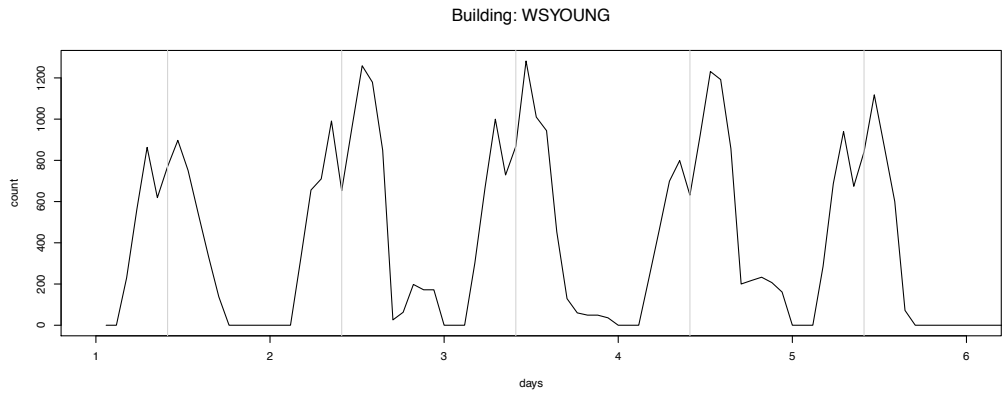
In R, we've created another data table (not dissimilar from the `arbuthnot` and `present` data sets you looked at in lab) that represents building occupancy (scheduled) over time...

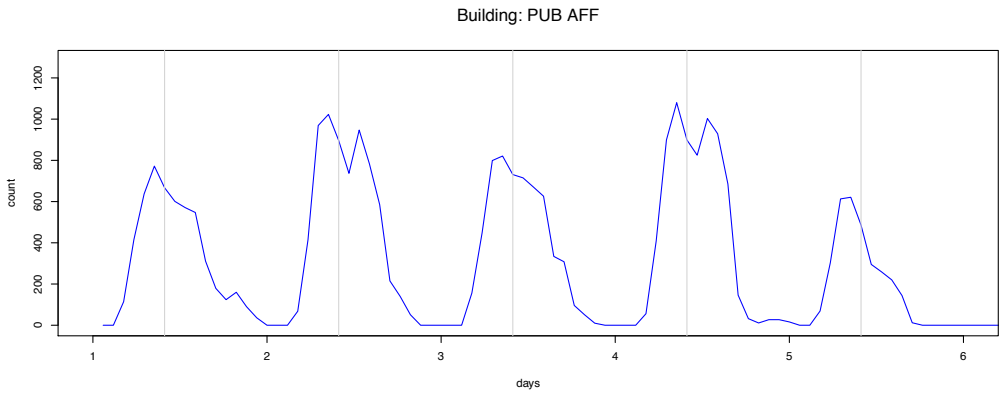
```
# subscripting below asks for the first 10 rows and 10 columns of buildings

> buildings[1:10,1:10]
    bnames 06:00M 07:00M 08:00M 09:00M 10:00M 11:00M 12:00M 13:00M 14:00M
1       AU      0      0      0      0      0      0      0      0      0
2  BIO SCI      0      0      0      0      0     59      0      0      0
3      BMC      0      0      0      0      0      0      0      0      0
4  BOELTER      0      0    300    356    698    567    674    664    627
5   BOTANY      0      0     30     90     54     54     30     60     30
6    BOYER      0      0      0      0      0      0      0      0      0
7    BROAD      0      0     48    307    313    448    385    281    372
8   BUNCHE      0      0     77    557    770    729    308    718    737
9  CAMPBEL      0      0      0     11     11      0      0      0      3
10 COLLINS      0      0      0      0      0      0      0      0      0

# the subscripting here asks for the last 10 rows and first 10 columns

> buildings[57:66,1:10]
    bnames 06:00M 07:00M 08:00M 09:00M 10:00M 11:00M 12:00M 13:00M 14:00M
57     SAC      0      0      9     29     35     16      0     22     22
58  SCULPT      0      0      0      0      0      0      0      0      0
59   SEMEL      0      0      0      0      0      0      0      1     11
60 SLICHTR      0      0      0      0      0      0     19     19     20
61     SMB      0      0      0    475    495    706    694     76     86
62  SPROUL      0      0      0      0      0      0      0      0      0
63 STRTHMR      0      0      0      0      0      0      0      0      0
64     TBA      0      0      0      0      0      0      0      0      0
65     UES      0      0      0      0      0      0      0      0      0
66 WGYOUNG      0      0    230    563    863    619    770    897    751
```

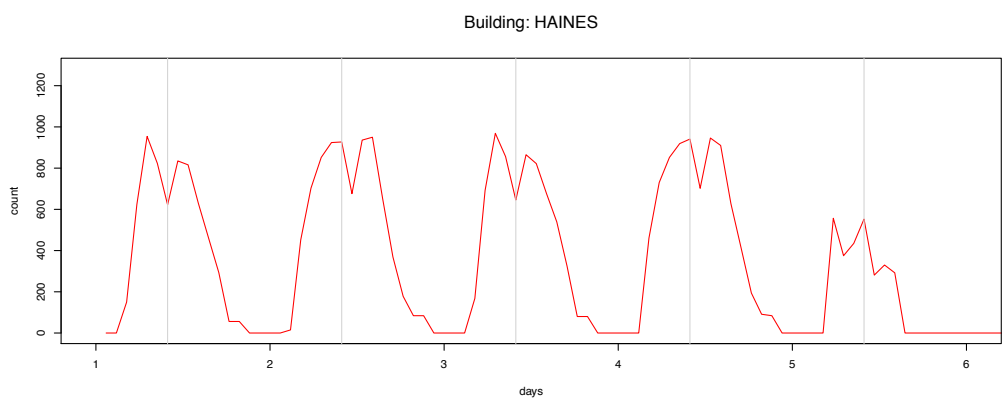
A building view

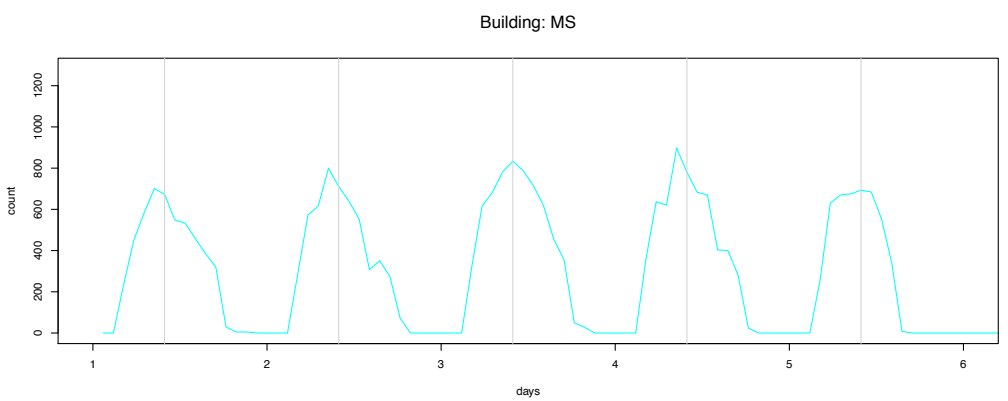
The data as we've defined them **deal entirely with occupancy** and the resulting table has 66 rows (one for each building the Registrar uses for classes) and 102 columns (occupancy is recorded for 17 hours between 6am and 10pm for 6 days for a total of 102 variables or columns)

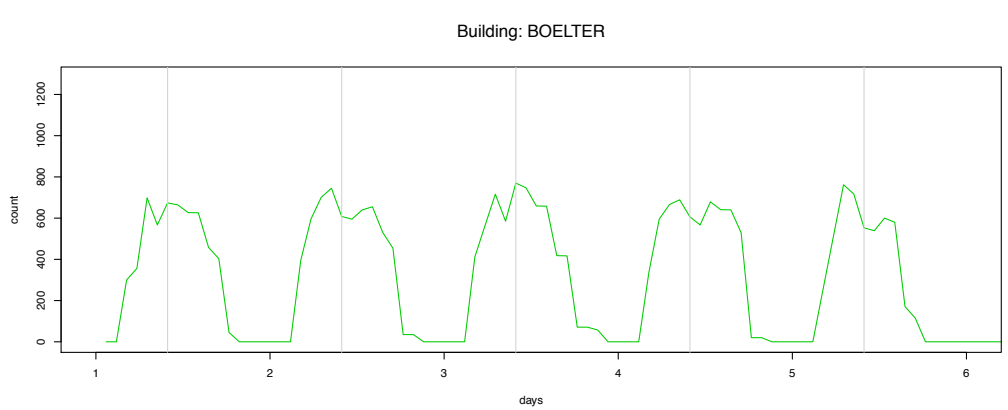
Given the time structure, we can make simple **time series plots** of occupancy during any week in Spring 2011 for the five most frequently used (as measured by total occupancy across the week) buildings -- Do you notice any differences?

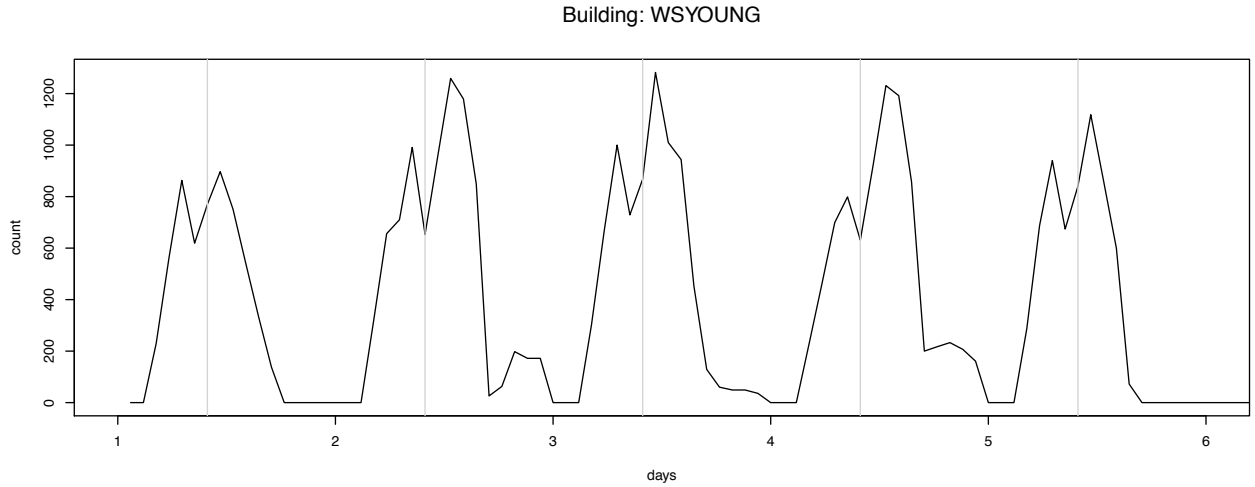

Building: WSYOUNG
count
0
200
400
600
800
1000
1200
1
2
3
4
5
6
days

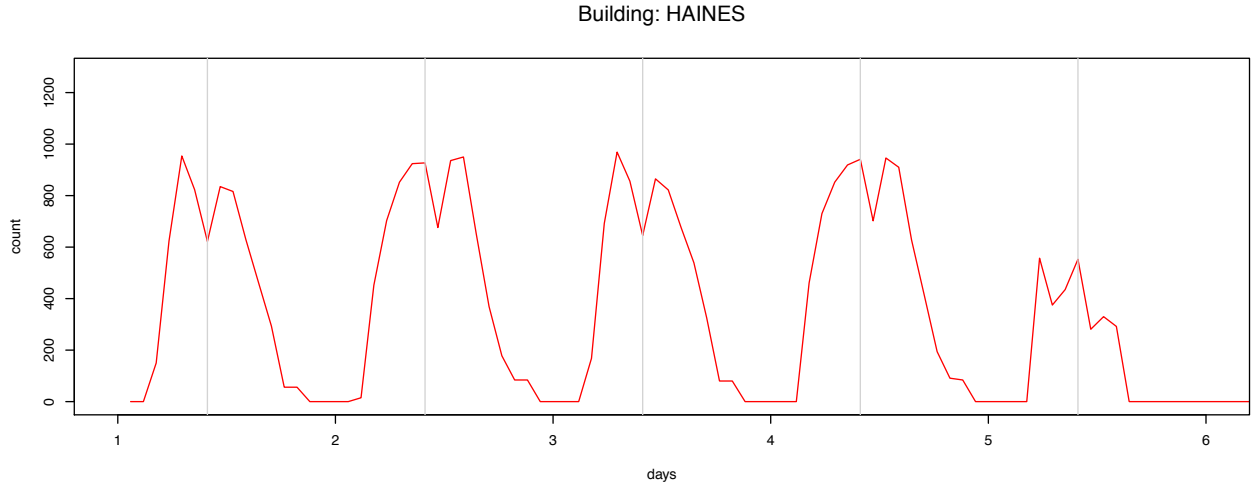

Building: PUB AFF
count
0
200
400
600
800
1000
1200
1
2
3
4
5
6
days

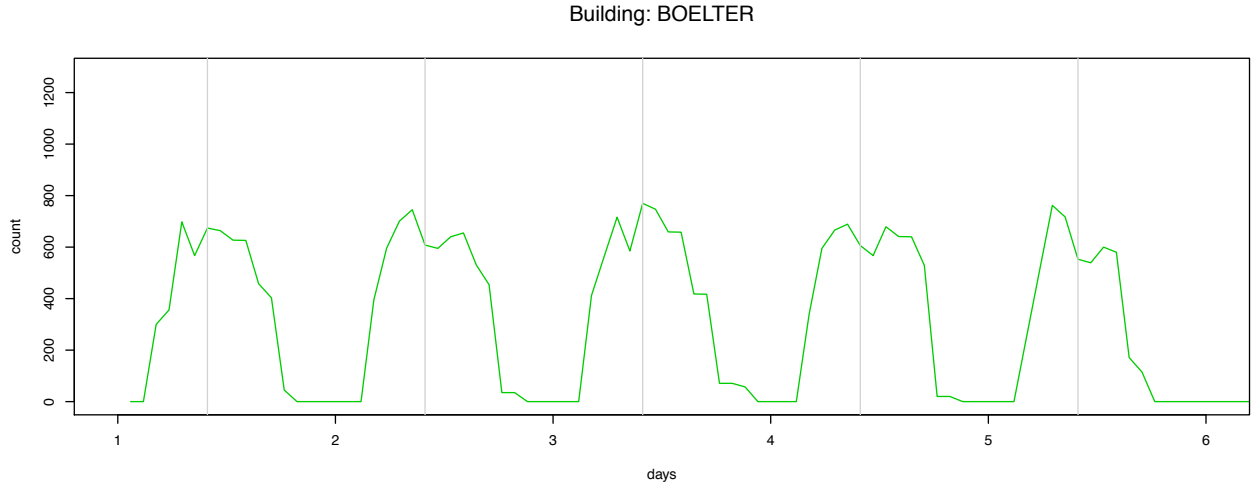

Building: HAINES
count
0
200
400
600
800
1000
1200
1
2
3
4
5
6
days

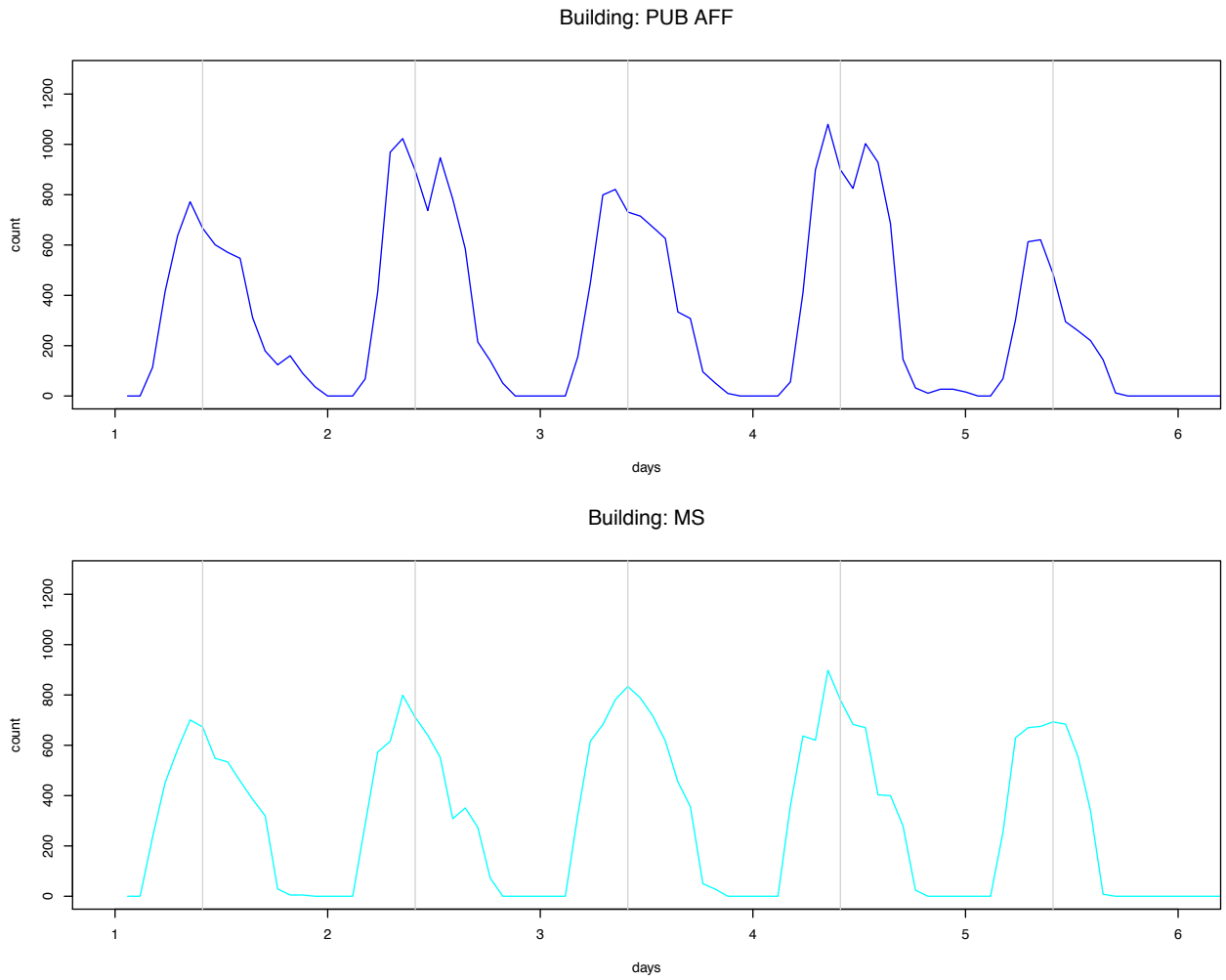

Building: MS
count
0
200
400
600
800
1000
1200
1
2
3
4
5
6
days


Building: BOELTER
count
0
200
400
600
800
1000
1200
1
2
3
4
5
6
days


Building: WSYOUNG
count
0
200
400
600
800
1000
1200
1
2
3
4
5
6
days


Building: HAINES
count
0
200
400
600
800
1000
1200
1
2
3
4
5
6
days

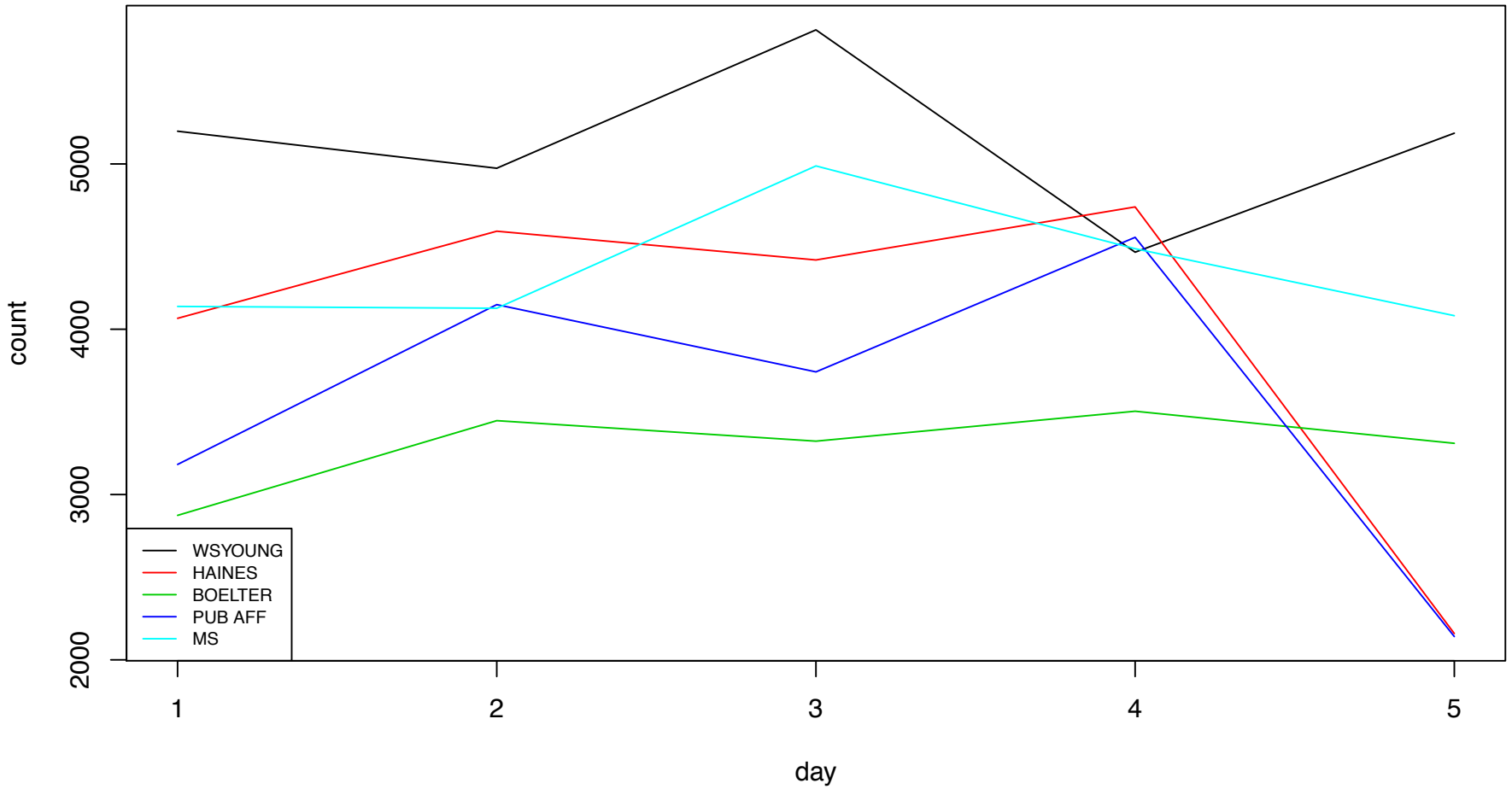

Building: BOELTER
count
0
200
400
600
800
1000
1200
1
2
3
4
5
6
days


Building: PUB AFF
count
0
200
400
600
800
1000
1200
1
2
3
4
5
6
days
Building: MS
count
0
200
400
600
800
1000
1200
1
2
3
4
5
6
days

A building view

We can aggregate across time as well, and examine **the daily breakdown of occupancy per day** -- That is, the number of students seen in each of these five buildings (we counted you twice if you had two classes in this room last quarter, for example)

The time series plot is simpler (as we choose to show daily counts for Monday through Friday) but maybe show the patterns of usage?


count
2000
3000
4000
5000
1
2
3
4
5
day
WSYOUNG
HAINES
BOELTER
PUB AFF
MS

A building view

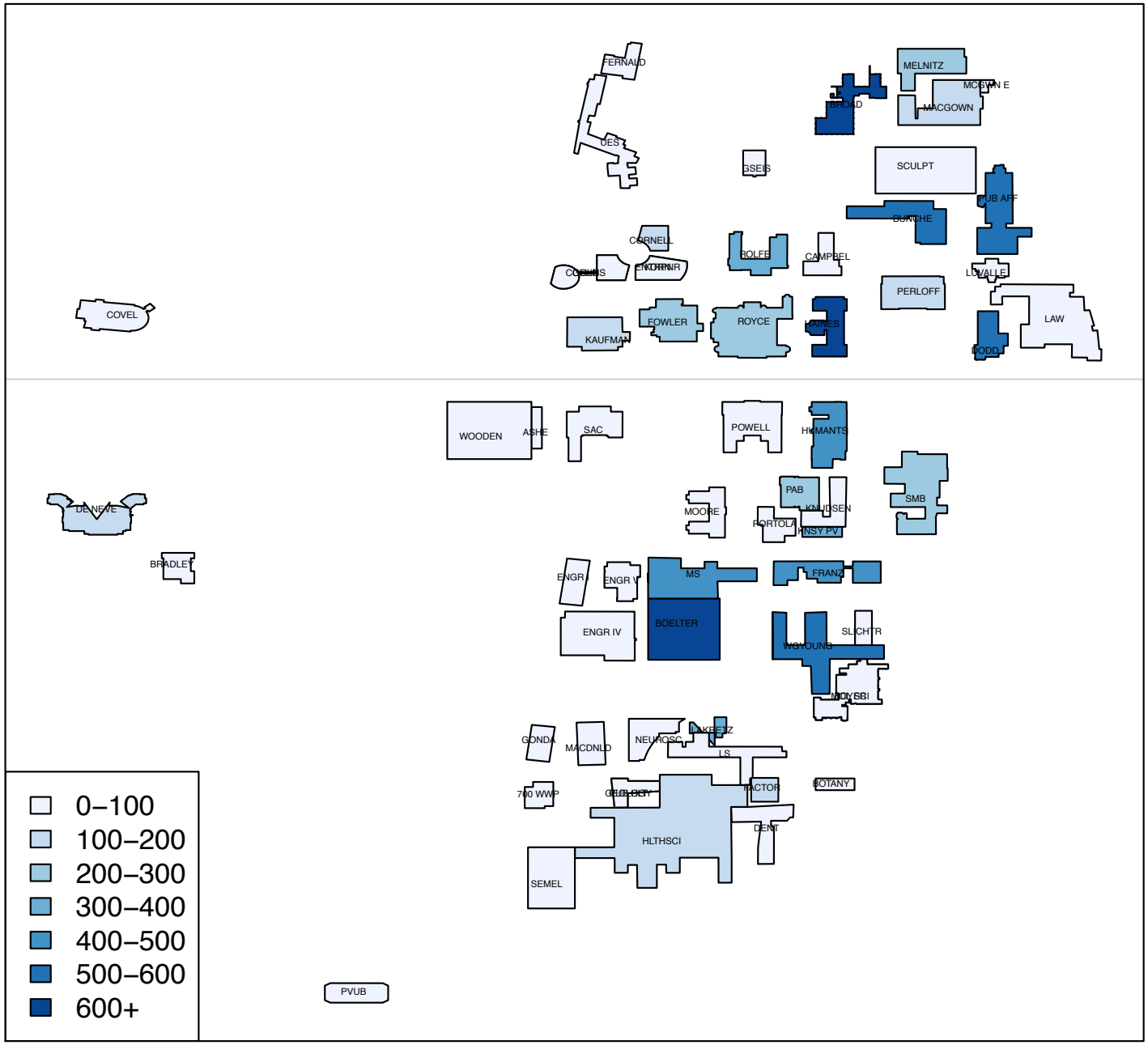
Over the quarter, will spend a fair bit of time with spatial data, but here we can at least give you a taste of what to expect -- We can **“join” our building occupancy data with a map of campus**

In this case, we have another kind of table, one in which the rows again refer to buildings but **the entries are spatial shapes** -- Polygons that describe the outlines of the buildings

A choropleth map shades regions according to some measurement or count or category (a variable from last lecture) and are often used to exhibit Census or other data -- Here we use occupancy...


FERNALD
MELNITZ
MCGWN E
BROAD
MACGOWN
UES
GSEIS
SCULPT
PUB AFF
BURCHE
CORNELL
ROLFE
CAMPBEL
LOVALLE
PERLOFF
COVEL
FOWLER
ROYCE
HAINES
LAW
KAUFMAN
DODD
WOODEN
ASHE
SAC
POWELL
HUMANTS
PAB
SMB
DENEVE
MOORE
KNUDSEN
FORTOLA
HNSY PV
BRADLEY
ENGR I
ENGR V
MS
FRANZ
ENGR IV
BOELTER
SLICHTR
WGYOUNG
GONDA
MACDNLD
NEUROSC
LAKRETZ
LS
700 WWP
FACTOR
BOTANY
DENT
HLTHSCI
SEMEL
PVUB

3:00 on a Monday Winter, 2011 (9,539 students)



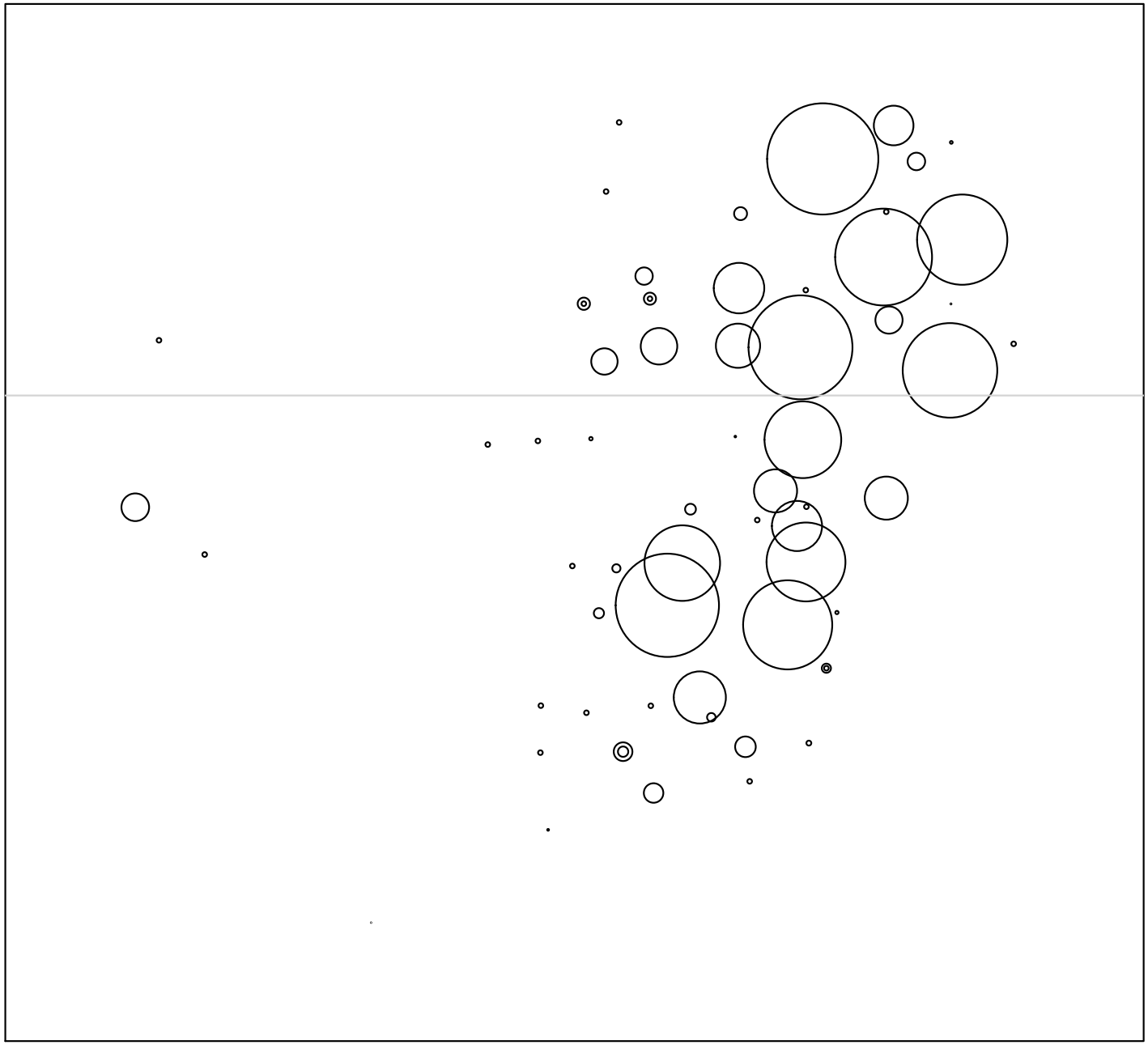
A building view

It is also common to render the values of a variable using the sizing of a geometric shape, say a circle -- **A bubble chart** renders (in this case) occupancy as a circle centered on the building's center with radius proportional to occupancy

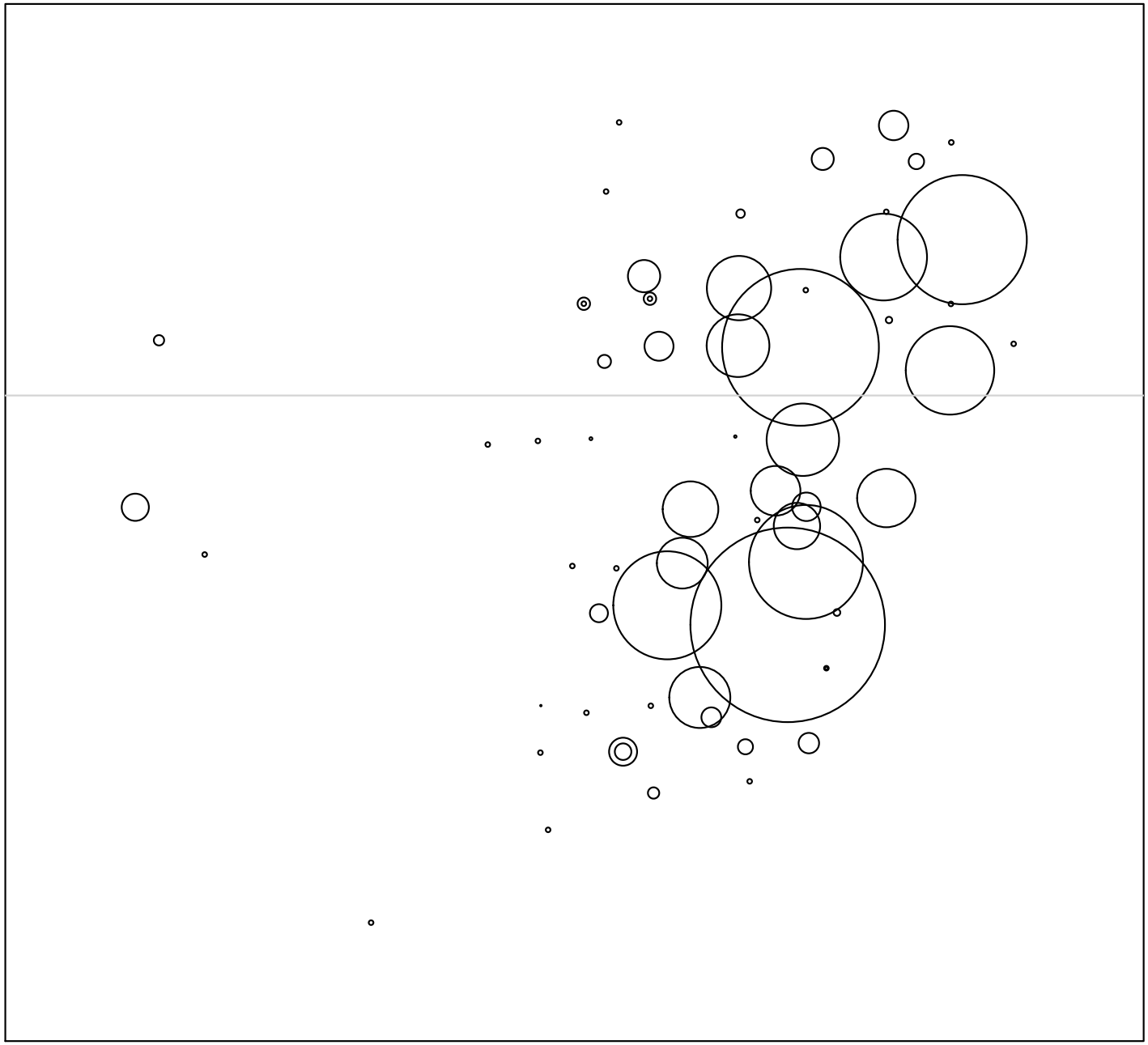
In R these are rendered using the function `symbols` and can be used to generate fairly complex shapes that might depend on more than just a single variable (say two values might go into the height and width of a box)

We'll have more fun with this kind of display later in the quarter -- For now, here's a few bubble charts of building occupancy at 15:00 last quarter on each weekday (the gray line runs between Royce and Powell)

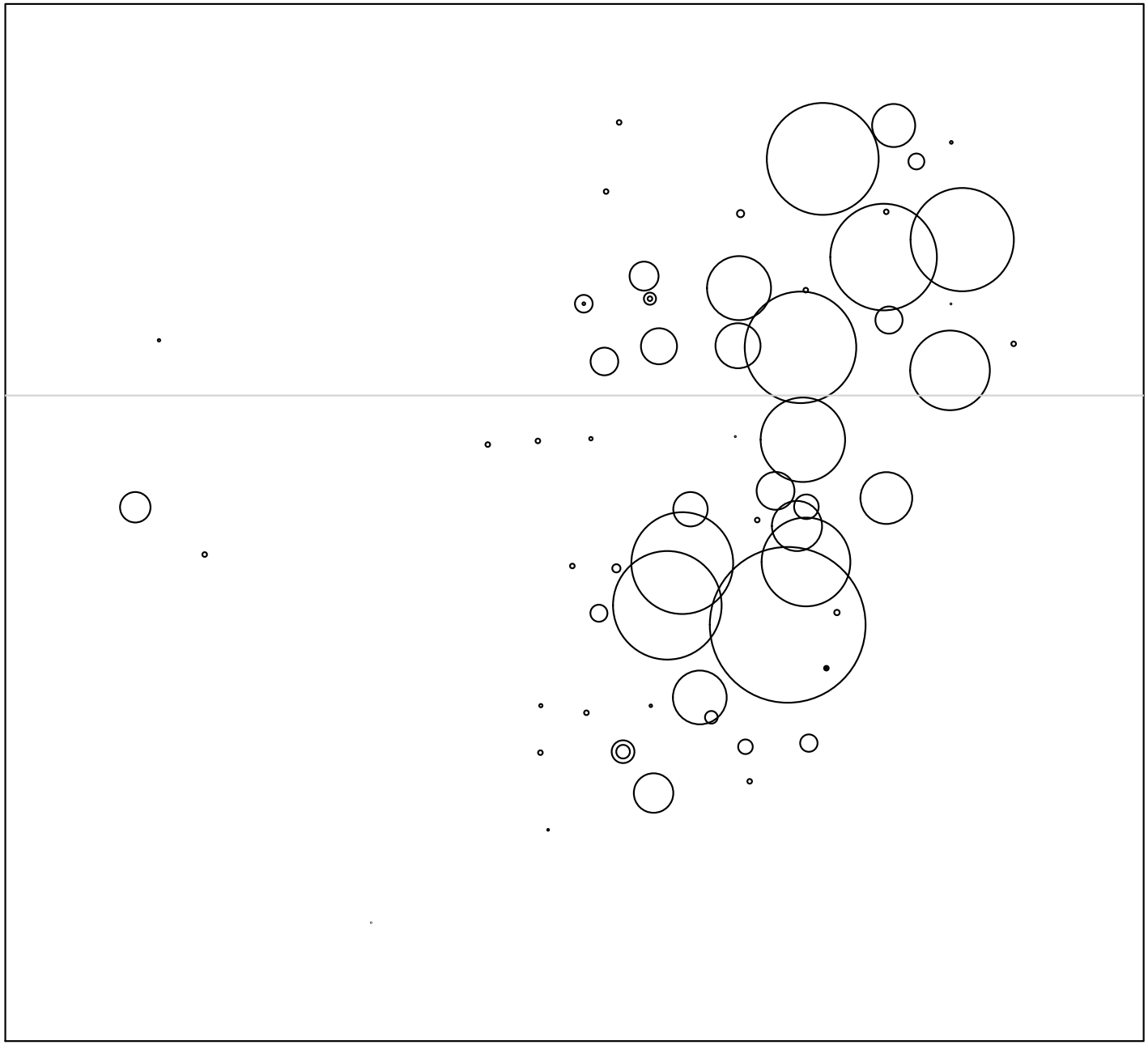
3:00 on a Monday Winter, 2011 (9,539 students)



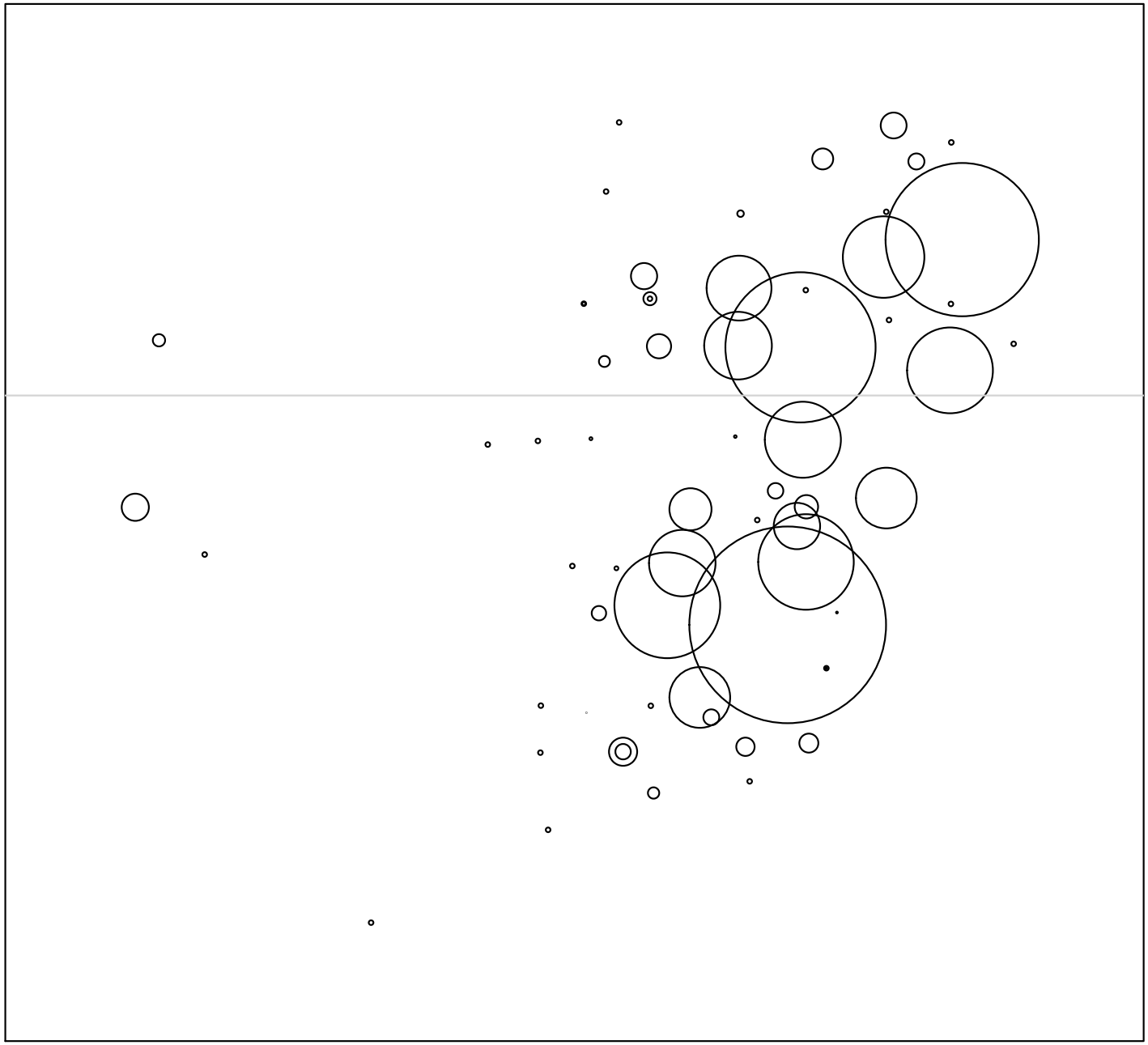
3:00 on a Tuesday Winter, 2011 (10,946 students)



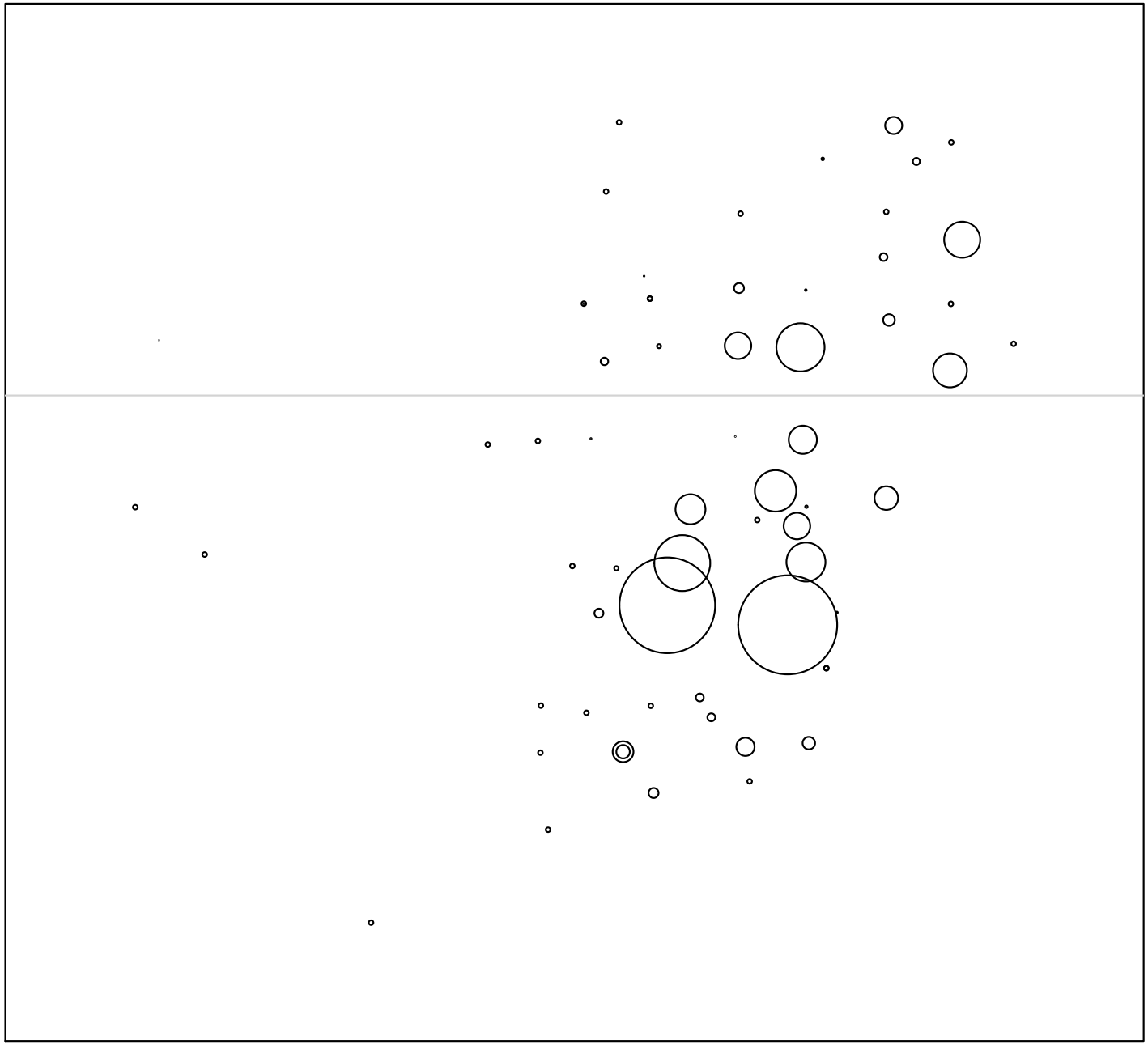
3:00 on a Wednesday Winter, 2011 (11,201 students)



3:00 on a Thursday Winter, 2011 (10,513 students)



3:00 on a Friday Winter, 2011 (4,667 students)



The registrar's data

In the last example, we combined two data sets -- One with the locations and shapes of the buildings on campus and the other consisting of hourly occupancy according to the registrar's course rosters

Joining data in this way is another fundamental concept that is worth talking about -- **Data are, well, promiscuous** in the sense that the seem to want (assigning agency to data seems like a problem) to be combined with other data

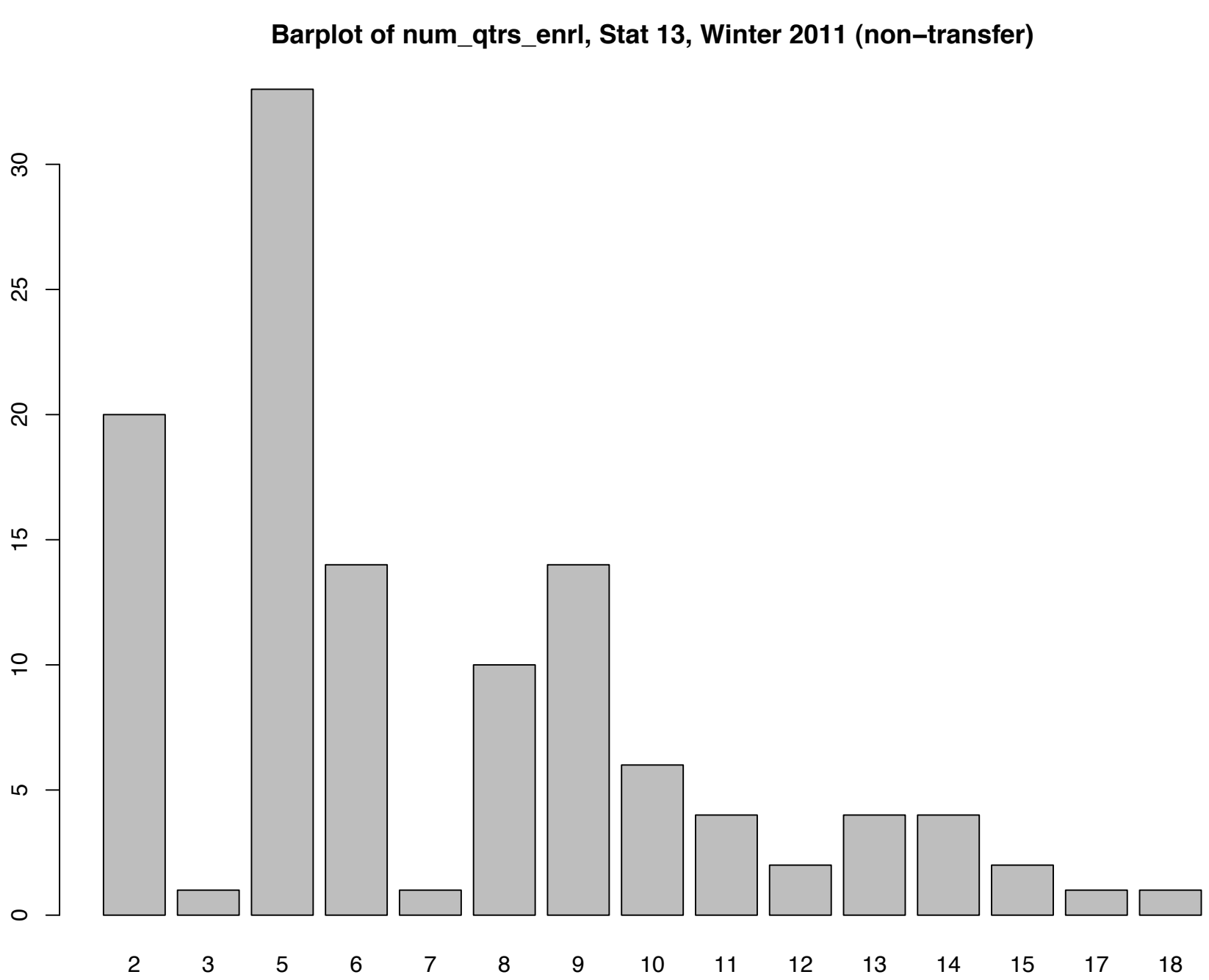
For example, we've been walking down a relatively innocent path counting students in classes or looking at characteristics of students' schedules last quarter -- Our $85 bought us a bit more data, however, that can be joined (using the student index) with our enrollment or student view...

```
> head(reg_student,n=30)
          id     level num_qtrs_enrl transfer_stu major_cd age sex ethnicity
1  816640632         U             2            N     0948  18   M        BL
2  816643648         G             2            N     0602  26   F        DS
3  816577472         U             1            N     0181  20   F
4  806029941         G            19            N     0249  23   F        JA
5  821748664         U            15            N     0928  22   F        MA
6  820969784         U            11            N     0303  22   M        VI
7  821030697         U            13            N     0181  22   M        WH
8  820915798         U            13            N     0111  22   M        WH
9  820850323         U            11            N     000G  21   F        CA
10 820902758         U            12            N     0440  21   M        WH
11 820899244         U            13            N     0303  21   M        CA
12 816656727         G             2            N     0249  22   F
13 806161843         U             5            N     0181  23   M        WH
14 821243947         G            17            N     0330  24   F        VI
15 821035130         U            19            N     0153  22   M        WH
16 821018740         U            14            N     0699  22   M        WH
17 821018200         U            14            N     0671  21   M        CA
18 820964972         U            13            N     0111  22   F        CA
19 820851318         U            12            N     0777  21   F        WH
20 820841575         G            11            N     0246  30   M        LA
21 820890154         U            12            N     0300  21   M        CA
22 816628294         U             2            Y     0345  28   M        WH
23 821228468         U            14            N     0439  24   M        LA
24 821965243         G            13            N     0330  23   M
25 821949318         G            17            Y     0393  33   F        WH
26 821747052         U            16            N     0294  22   M        WH
27 821030837         U            12            N     0090  21   F        OT
28 820916332         U            13            N     0345  21   F        EI
29 820912547         U            12            N     0181  21   F        WH
30 820968604         U            12            N     0330  22   M        TH
```

Registrar's data

Joining these more demographic variables with the enrollment event or student data can lead us to **ask an expanded set of questions** -- And this is how the enterprise progresses

We'll have more to say about the basic operations or behaviors (in Latanya Sweeny's terminology) around data over the course of the quarter -- For now, we can compare examine some simple statistics about last term's Statistics 13 cohort


Barplot of num_qtrs_enrl, Stat 13, Winter 2011 (non−transfer)
0
5
10
15
20
25
30
2
3
5
6
7
8
9
10
11
12
13
14
15
17
18

Having tied up our loose ends...

We'll now return to **the exploratory data analysis we had been pursuing** in the last lecture -- We've seen barplots and histograms as tools to exhibit the "shape" of a data set (or the distribution of a variable)

Today we'll continue with this lesson, providing you with enough material to conduct your own investigation in Lab on Thursday -- By way of a roadmap, next lecture we will begin to examine **formal hypothesis testing via a permutation or rerandomization procedure** (our test case will be clinical trials)

For now, however, some new data...

The BRFSS

The Behavioral Risk Factor Surveillance System is **the world's largest telephone survey** and it is designed to track health risks in the United States -- Like many surveys, the BRFSS works with only a sample of a larger population

With over 200 million adults in the United States, the CDC couldn't possibly contact their entire population* -- Instead, they selected around 350 thousand adults, calling tends of thousands of people per month

* Even if time wasn't an issue, is it possible to contact every adult in the U.S.?

Turning Information Into Public Health

The Behavioral Risk Factor Surveillance System (BRFSS) is a state-based system of health surveys that collects information on health risk behaviors, preventive health practices, and health care access primarily related to chronic disease and injury. For many states, the BRFSS is the only available source of timely, accurate data on health-related behaviors.

BRFSS was established in 1984 by the Centers for Disease Control and Prevention (CDC); currently data are collected monthly in all 50 states, the District of Columbia, Puerto Rico, the U.S. Virgin Islands, and Guam. More than 350,000 adults are interviewed each year, making the BRFSS the largest telephone health survey in the world. States use BRFSS data to identify emerging health problems, establish and track health objectives, and develop and evaluate public health policies and programs. Many states also use BRFSS data to support health-related legislative efforts.

The bigger picture

There is an implicit hope that the sample of adults identified by the CDC is in some way representative of the larger population within the United States -- **If it is, we can begin to infer aspects of the population from the sample**

We do this at least informally every time the press reports the President's approval rating or we hear about the success rate of a new AIDS treatment

Statistical inference is the process of drawing conclusions about a population, based on a observations in **a sample from that population** -- As we mentioned last time, **modern inference often involves various phases of exploratory data analysis**

Here, numerical and graphical descriptions of the data are used to help us uncover patterns, to get a sense of what the data look like

Our data

The BRFSS consists of responses from hundreds of thousands of people -- In this lecture and in your lab, we will only look at a subset (another sample, if you will) of 20 thousand people (although they will be two different samples, just for grins)

Here are the first ten responses in our data set -- Each row refers to a different person and each column to their response to a given question

	state	genhlth	physhlth	exerany	hlthplan	smoke100	height	weight	wtdesire	age	gender	sprawl
1	22	good	0	0	1	0	70	175	175	77	m	77.27268
2	25	good	30	0	1	1	64	125	115	33	f	45.72318
3	6	good	2	1	1	1	60	105	105	49	f	48.73611
4	6	good	0	1	1	0	66	132	124	42	f	14.21793
5	39	very good	0	0	1	0	61	150	130	55	f	61.64302
6	42	very good	0	1	1	0	64	114	114	55	f	57.74011
7	6	very good	0	1	1	0	71	194	185	31	m	48.73611
8	48	very good	1	0	1	0	67	170	160	45	m	45.03769
9	6	good	2	0	1	1	65	150	130	27	f	32.24949
10	48	good	3	1	1	0	70	180	170	44	m	45.87459

Variables

`state`
Where does the respondent live?

`genhlth`
Respondents were asked to evaluate their general health values are excellent, very good, good, fair, poor

`physhlth`
The number of days out of the last 30 that the respondent was in poor health

`exerany`
1 if the respondent exercised in the last month and 0 otherwise

`hlthplan`
1 if the respondent has some form of health coverage and 0 else

`smoke100`
1 if the respondent has smoked at least 100 cigarettes in their entire life and 0 otherwise

Variables

`height` in inches

`weight` in pounds

`wtdesired` desired weight in pounds

`age` in years

`gender` labeled “f” and “m”

`sprawl`

a variable not originally included in the BRFSS, but added by a research at JHU; it ranges from 0-100, with low values indicating densely populated regions and high values indicating urban sprawl (New York City = 6.7, L.A = 10.6, and Atlanta = 80.7)

Making things a little formal: Samples

A sample is a collection of persons or things on which we measure one or more variables; the number of items in the collection is known as **the sample size**

When you encounter a data set that represents a sample of some larger population, you should again be ready to ask questions: How did these people or things come to be in the sample? What were the motivations and incentives for the people or organizations making the selections?

Our ability to reason from a sample to a larger population could be affected by the way in which a sample was formed; but more on this later...

Surveys

Your text (in Chapters 2 and 3) spends a good deal of time illustrating the various designs or patterns researchers follow when conducting experiments or conducting surveys

The BRFSS is a **telephone survey** of adults -- The interviewer must first make sure that the number they are calling is a residence, then assess how many adults live at the residence and if greater than one, attempts to speak to the person with the most recent birthday

The BRFSS has started to trial **mobile phone sampling** as well, with the obvious constraints that it has to be a safe time for the interviewee to answer questions (and if not, the interviewer calls back at a more appropriate time)

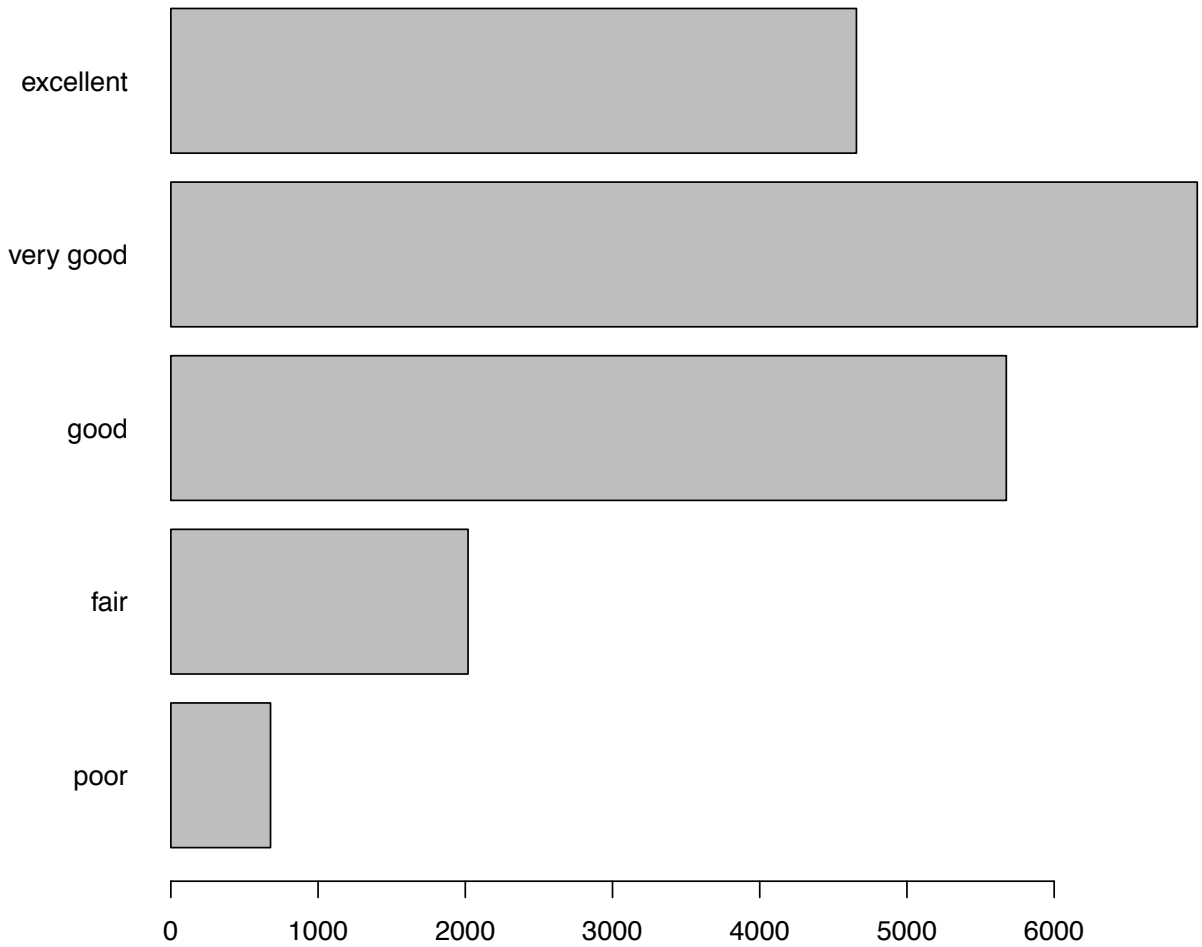
At the core, however, the BRFSS is attempting to generate **a random sample of the U.S. adult population** -- Next time we'll talk about how a computer can be used to generate random phenomena like simulating draws from a hat

Tables and barplots

For qualitative variables or discrete quantitative variables with a small number of values, we saw that simple tables and graphical displays computed from those tables gave us a pretty effective view into the data

Here are simple tabular displays of some of the qualitative variables in our data set -- What do we notice?

```
> table(cdc$gender)

    m     f
 9569 10431

> table(cdc$exerany)

    0     1
 5086 14914

> table(cdc$smoke100)

    0     1
10559  9441

> table(cdc$genhlth)

excellent very good      good      fair      poor
     4657      6972      5675      2019       677

> barplot(table(cdc$genhlth))
```

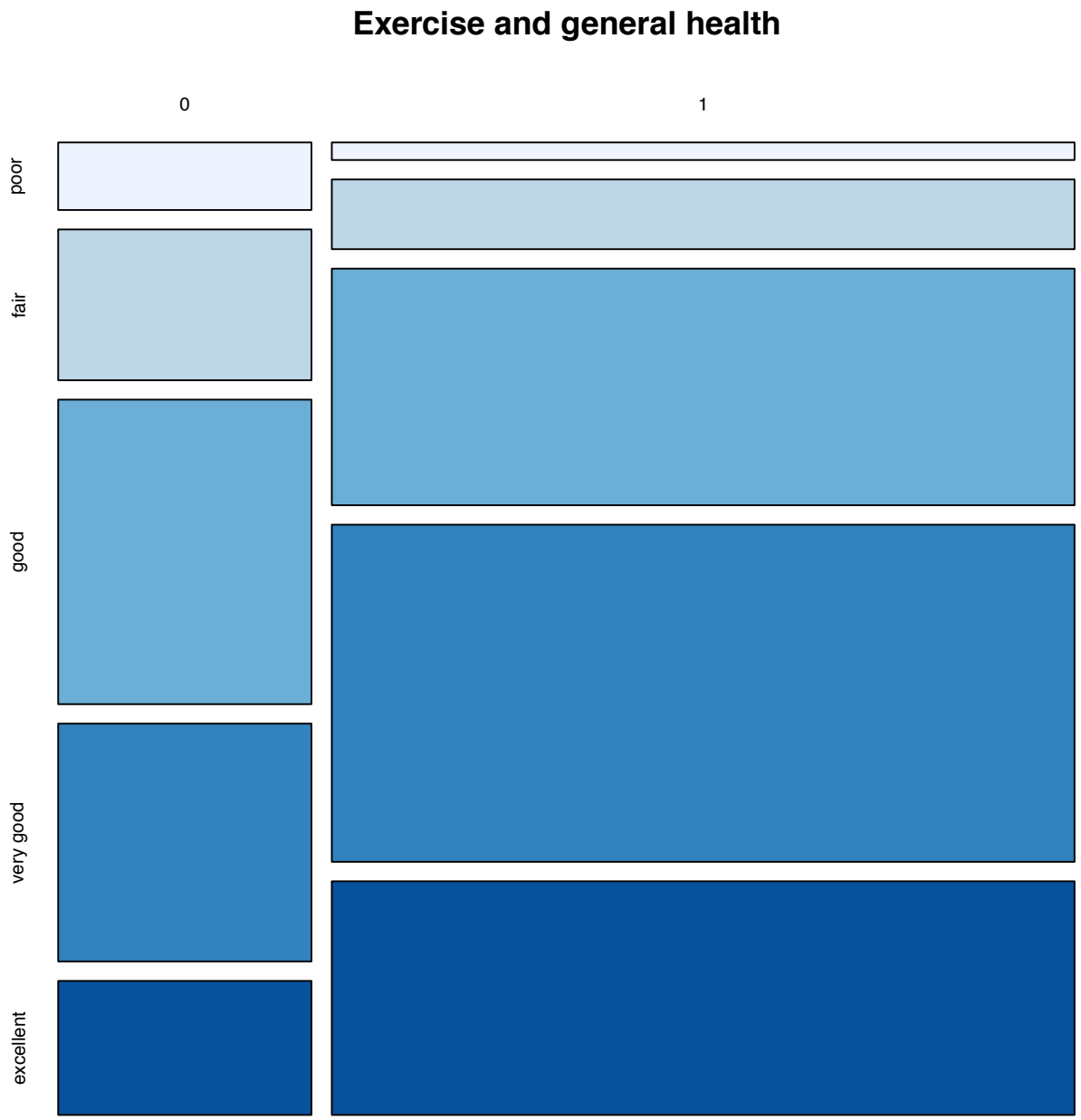

excellent
very good
good
fair
poor
0
1000
2000
3000
4000
5000
6000

Questions

While these one-dimensional summaries are interesting, they can't address certain questions we might bring to the data -- For example, does exercise have any effect on what people feel about their general health?

For this, we might consider a two-way table and an associated mosaic plot -- What do you see?

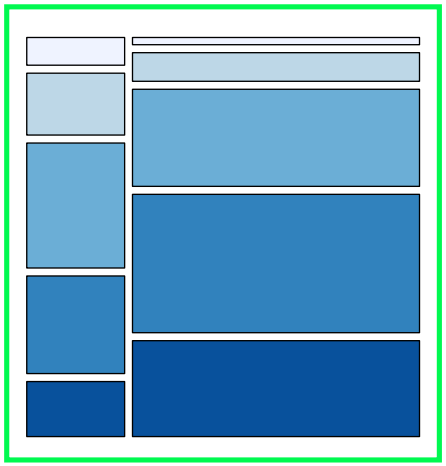
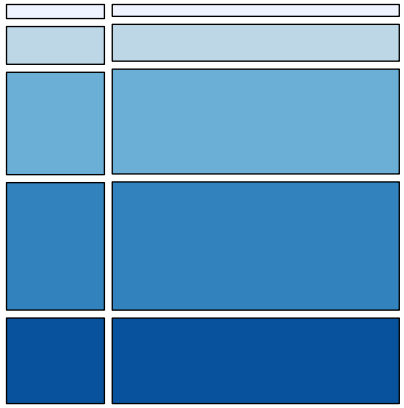
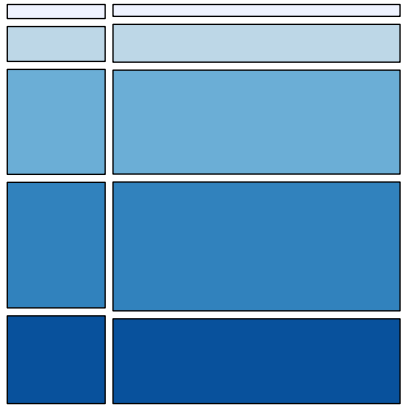
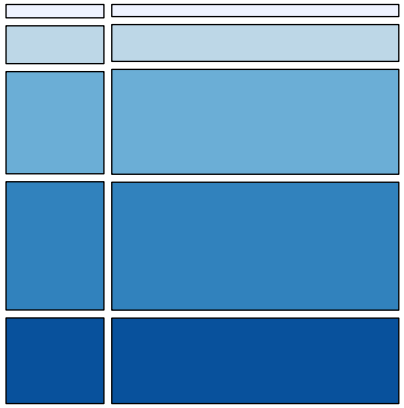
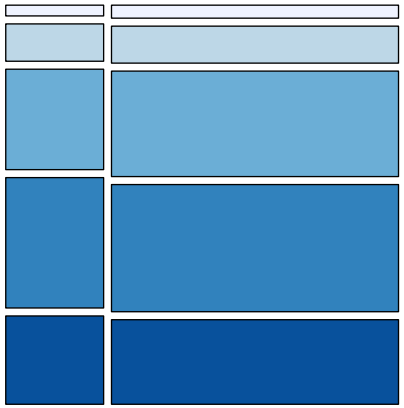
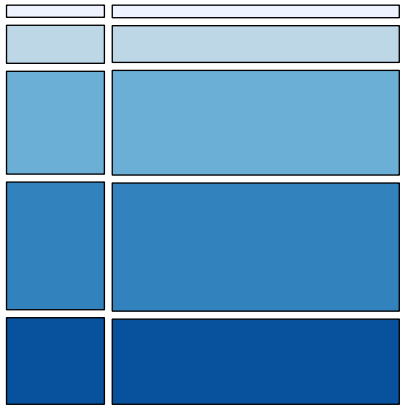
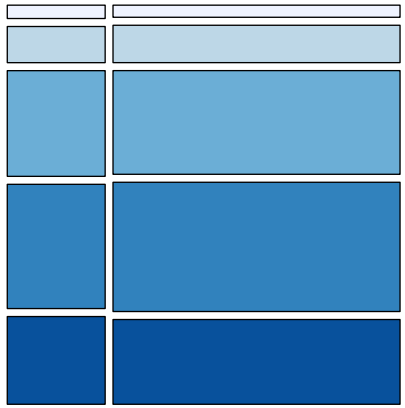
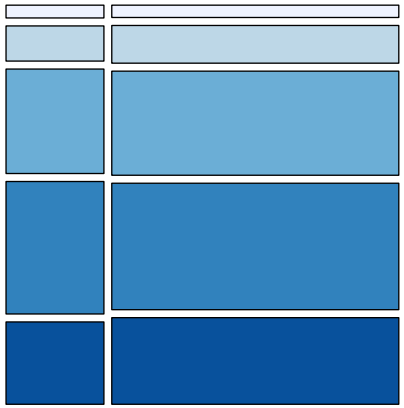
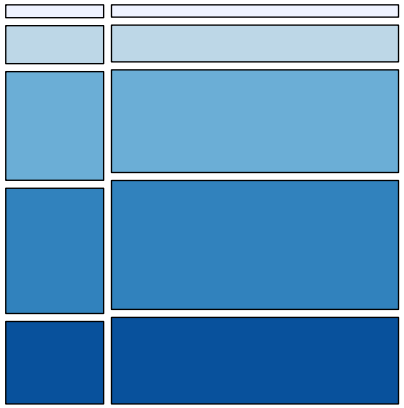
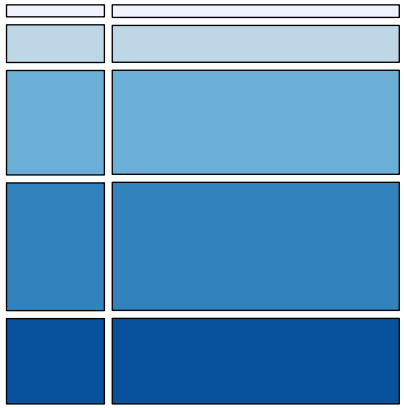
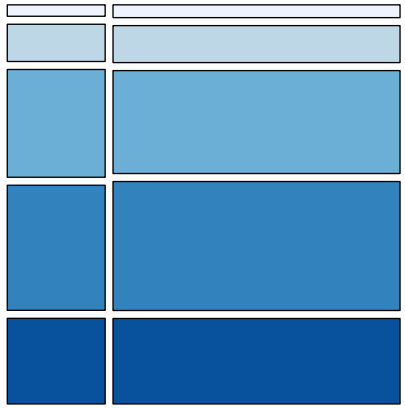
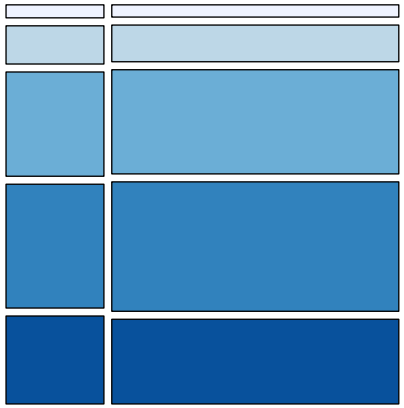
```
> table(cdc$exerany,cdc$genhlth)

    poor fair good very good excellent
  0  384  857 1731      1352       762
  1  293 1162 3944      5620      3895

# and now the plot -- we pass two arguments, the table
# we want to visualize and a title for the plot

> mosaicplot(table(cdc$exerany,cdc$genhlth),
             main="Exercise and general health")
```


Exercise and general health
0
1
poor
fair
good
very good
excellent

Association or not?

There seems to be some association here between exercise and general health -- That makes intuitive sense but you're relative newbs when it comes to "area" plots, it's worthwhile asking **what "no association" would look like**

Any ideas? How might we use R to simulate that for us?

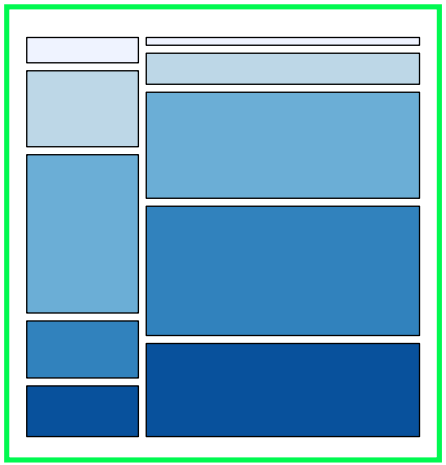
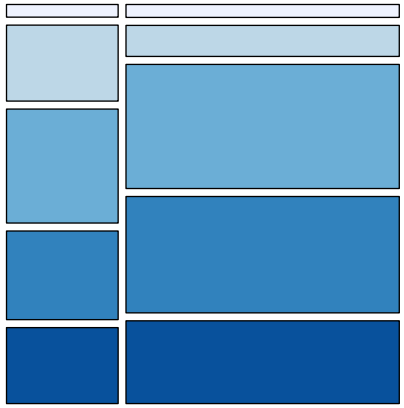
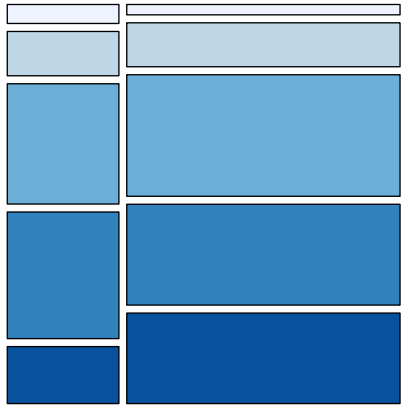
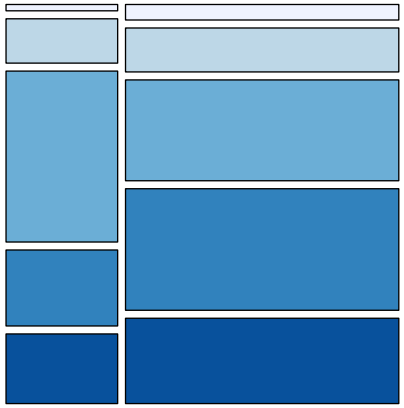
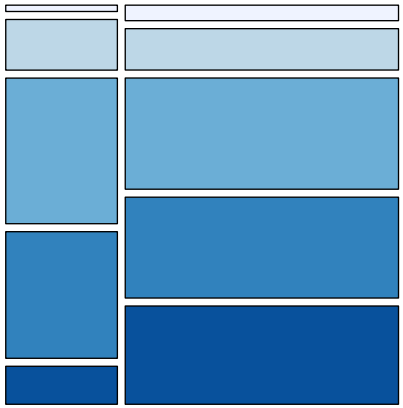
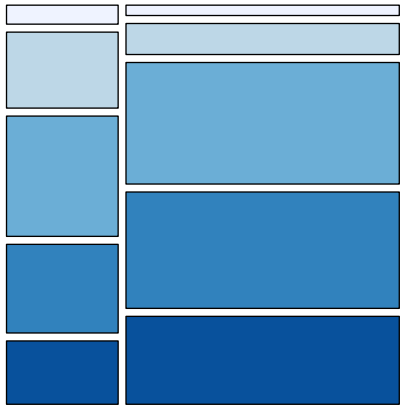
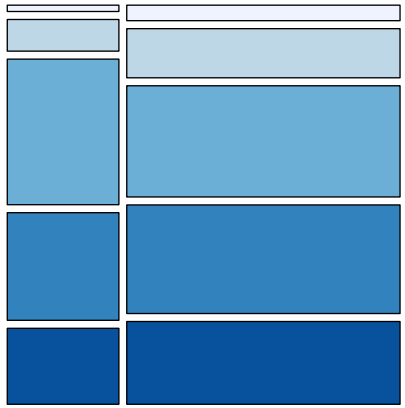
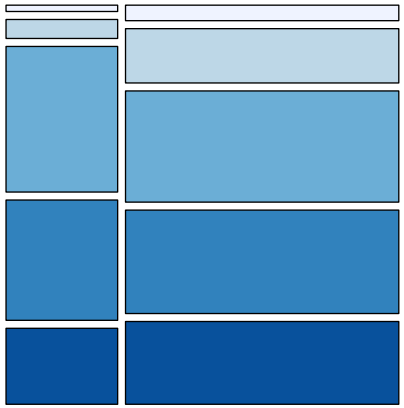
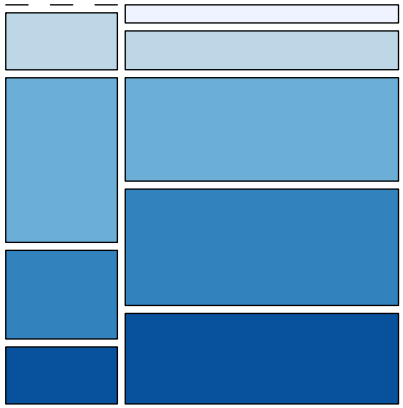
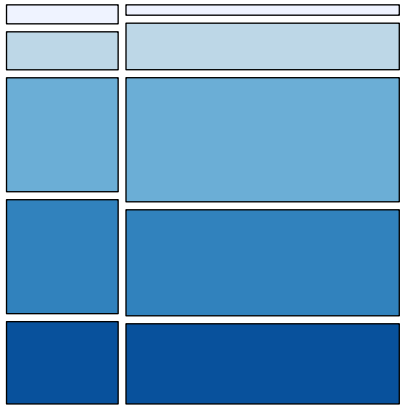
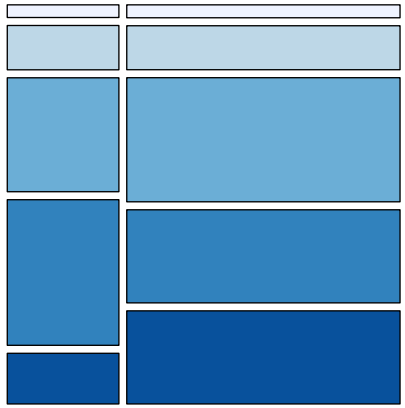
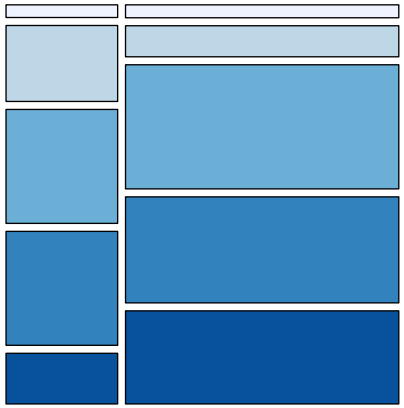













Simulation from a null model

By permuting the data associated with each person, we are generating data that, by definition, **should exhibit no pattern of association** -- For the moment we are using this device to help us calibrate our eyes so that we understand the tool we're working with

Notice that **there isn't much variability in the plots that we simulated because our sample size is so large** that by permuting things, we tend to get pretty stable proportions in each category -- This will change dramatically with smaller sample sizes

In the next lecture, we'll see that plots like these can also be used as a kind of visual hypothesis test (sexy!)

Creating new variables

BMI (Body Mass Index) is defined to be

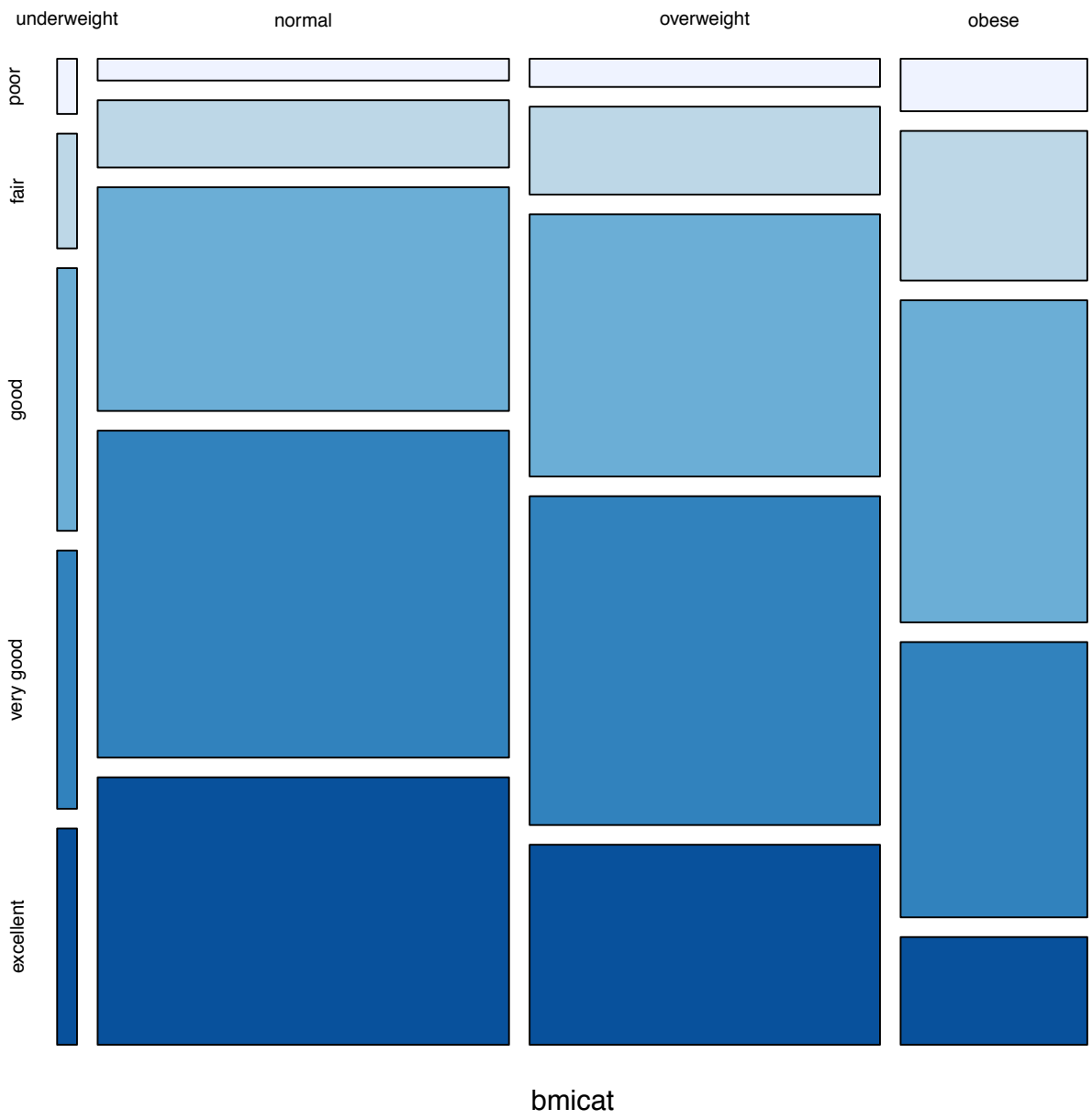
$$\text{BMI} = 703 \times \frac{\text{weight in pounds}}{(\text{height in inches})^2}$$

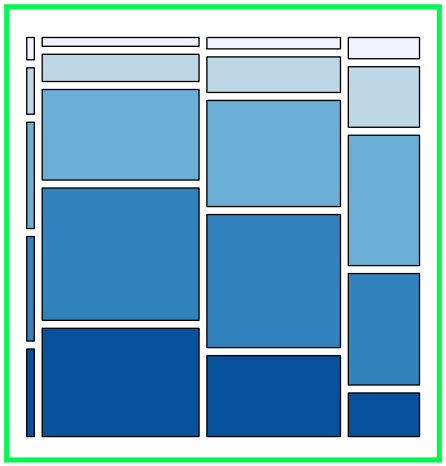
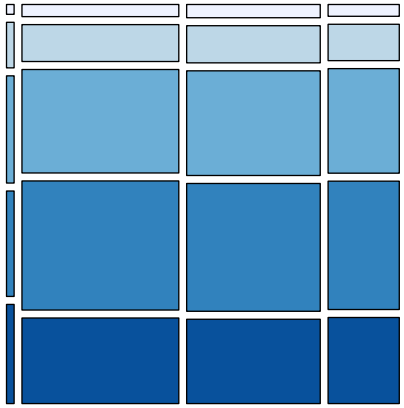
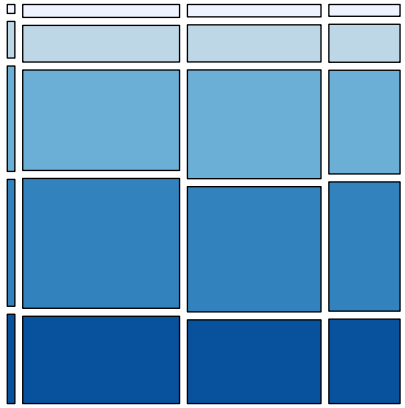
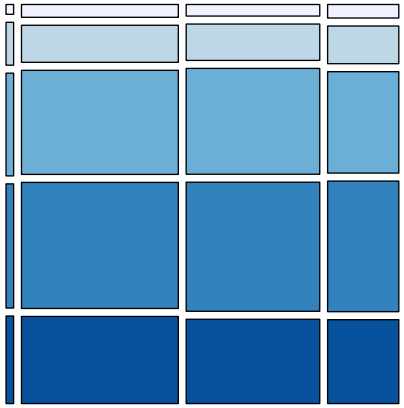
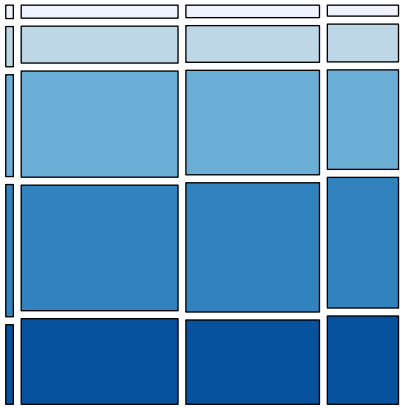
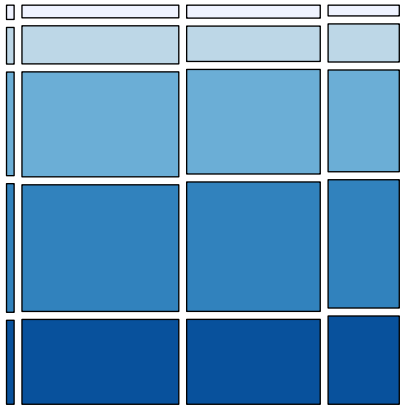
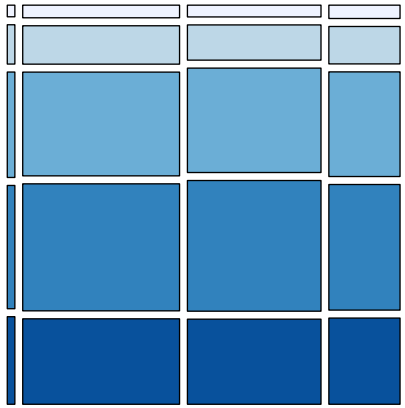
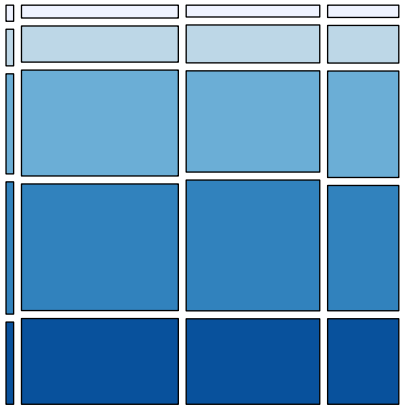
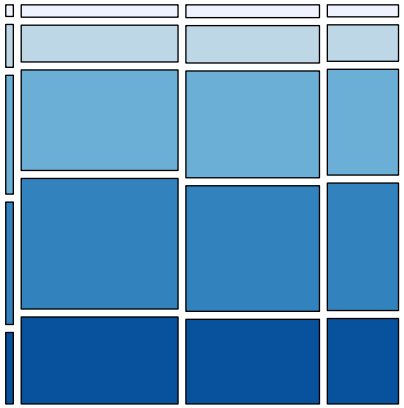
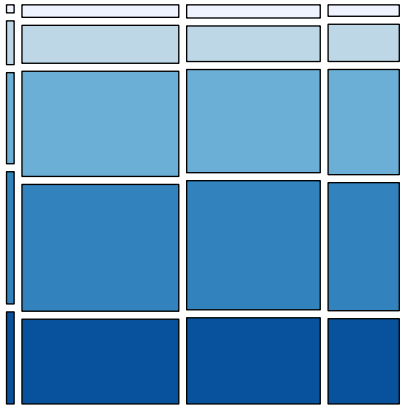
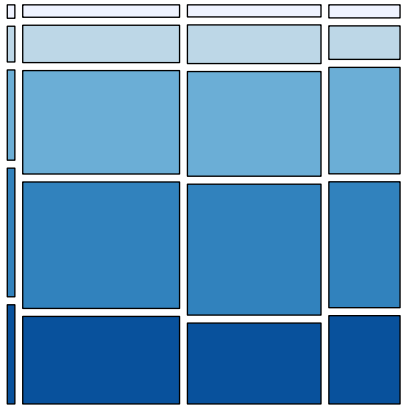
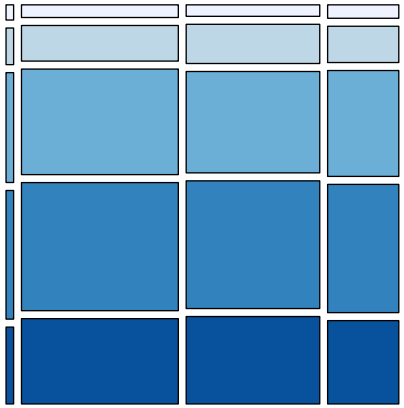
We can derive this from our data set and create a new quantitative variables

The CDC interprets these limits as follows

BMI	**Weight Status**
Below 18.5	Underweight
18.5 – 24.9	Normal
25.0 – 29.9	Overweight
30.0 and Above	Obese

```
> bmi <- 703*cdc$weight/(cdc$height)^2

# summary() is a handy function that will return a summary of the
# object you feed it -- Here we get a 6-number summary of the bmi values

> summary(bmi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.40   22.71   25.60   26.31   28.89   73.09

# now carve them into levels (this is advanced so ignore it on first reading)

> bmicat <- cut(bmi,c(0,18.5,25,30,100))
> levels(bmicat) <- c("underweight","normal","overweight","obese")

# and now summary gives you a table of the different category counts

> summary(bmicat)
underweight      normal  overweight       obese
        411        8496        7237        3856

# barplot!

> barplot(table(bmicat),horiz=T,main="BMI categories")

# and a mosaic plot!

> table(bmicat,cdc$genhlth)

bmicat        poor fair good very good excellent
  underweight   25   52  119       117        98
  normal       204  630 2095      3062      2505
  overweight   225  701 2092      2623      1596
  obese        223  636 1369      1170       458

> mosaicplot(table(bmicat,cdc$genhlth), main="BMI and general health")
```


BMI and general health
underweight
normal
overweight
obese
poor
fair
good
very good
excellent
bmicat

Histograms

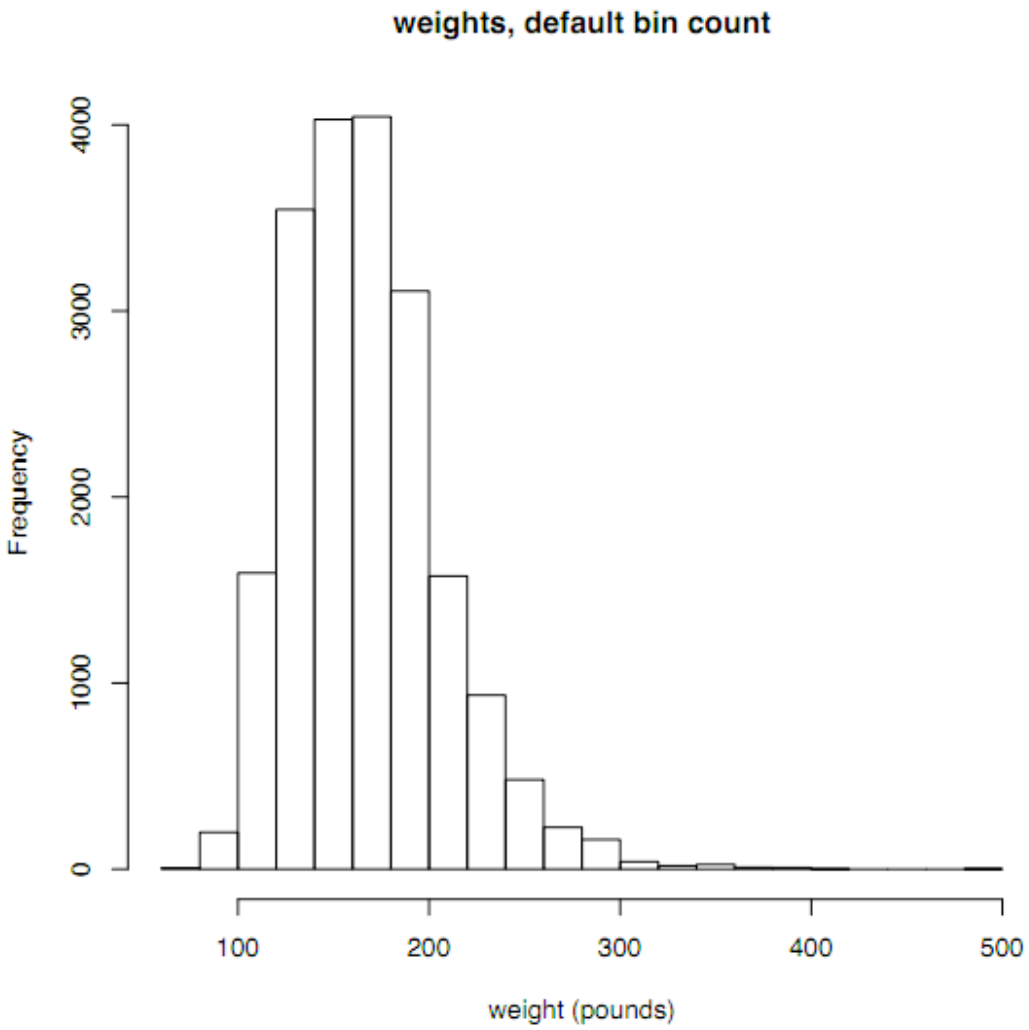
As we noted last time, a histogram groups or bins the data and, like a barplot, presents the number of data points that fall into each group

This display involves a “tuning parameter”; that is, we are free to choose how many bins we want to make the display

In situations like this, it is always good to vary the number of bins and examine the plot for any structure that emerges; in so doing, we want to get a sense of the “shape” of the data

```
> hist(cdc$weight)
```

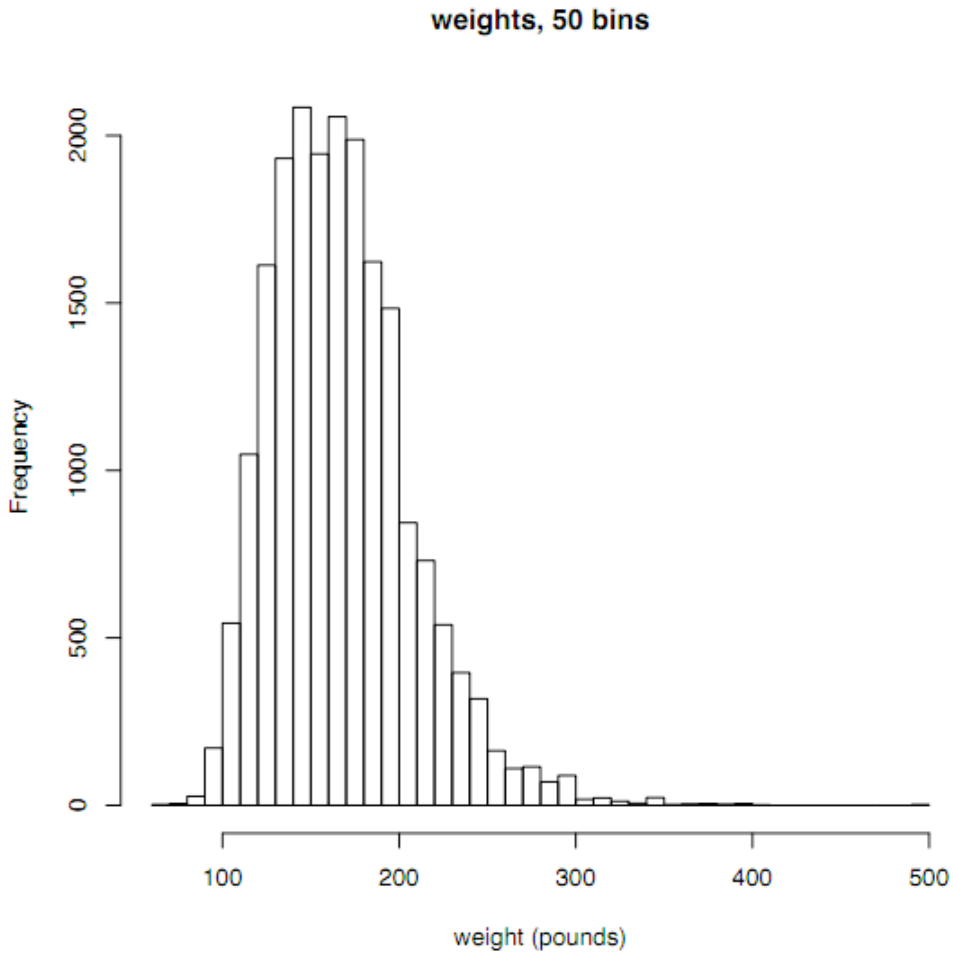
Given our discussion from last lecture, what can you say about the shape of this distribution?

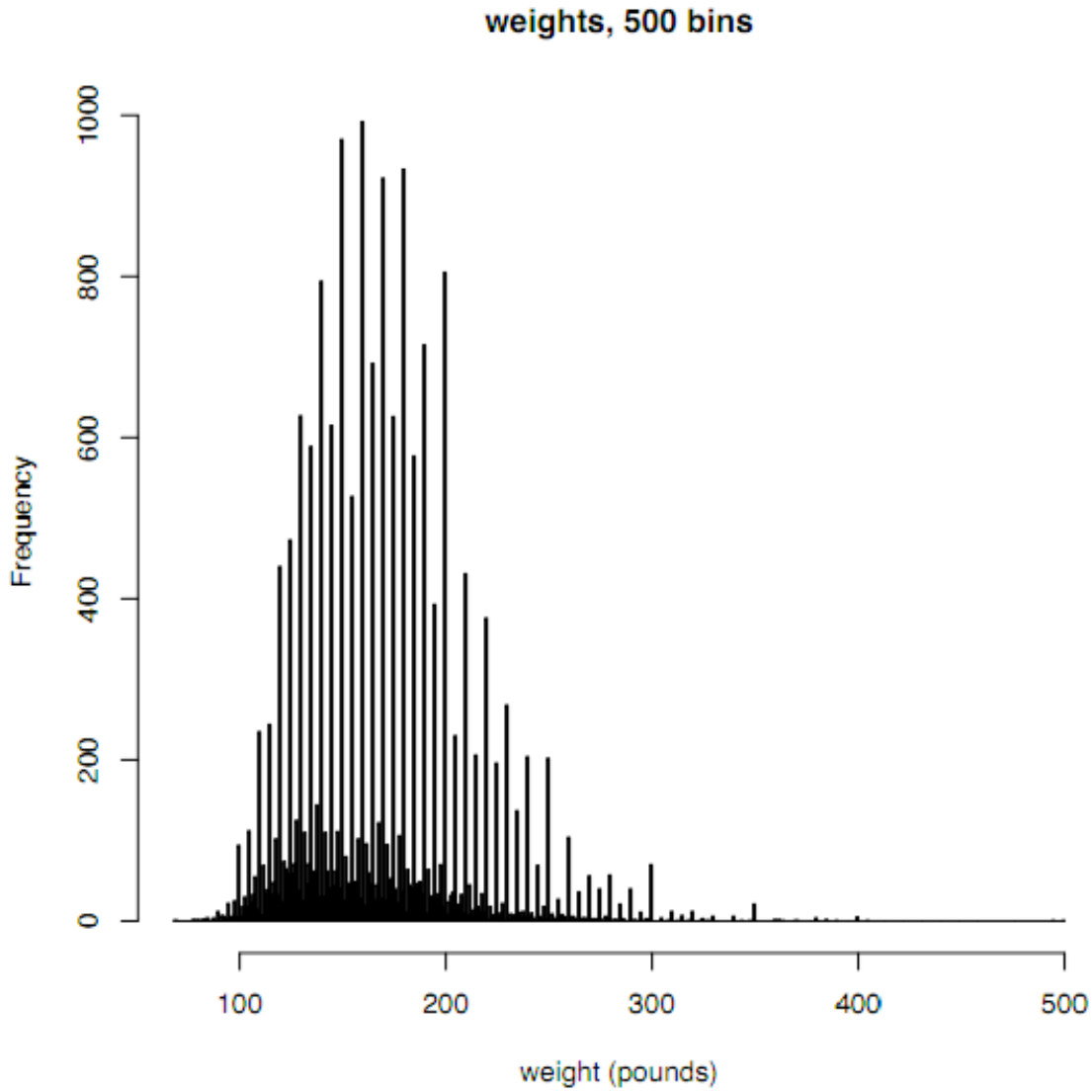

weights, default bin count
Frequency
0
1000
2000
3000
4000
100
200
300
400
500
weight (pounds)

And, varying bin widths...

With R, varying the size of our bins is as easy as

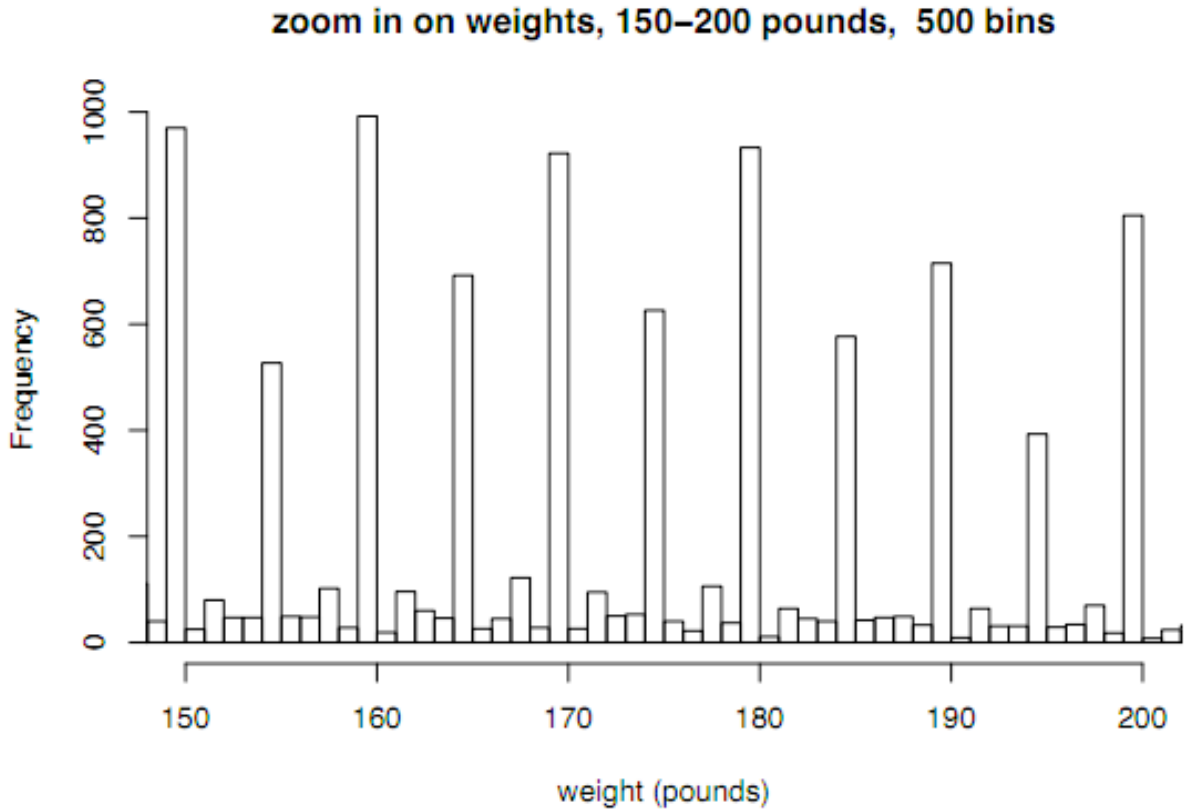
```
> hist(cdc$weight,breaks=50)
> hist(cdc$weight,breaks=500)
```


weights, 50 bins
Frequency
0
500
1000
1500
2000
100
200
300
400
500
weight (pounds)


weights, 500 bins
Frequency
0
200
400
600
800
1000
100
200
300
400
500
weight (pounds)

Varying bin sizes

By changing the bin size, we can uncover features in the data; in this case we uncover a basic fact about how people report their weights



Default bin size

It is often the case that we don't want to think very hard about how many bins or groups to use when drawing a histogram; the hist() function in R uses a rule of thumb for setting the number of bins based on our sample size

$$\text{number of bins} \approx \log_2(n) + 1$$

Where might a rule like this come from?

Default bin size

At the moment, we are using histograms as a tool to investigate **the structure of data, of a quantitative variable**

We often use histograms, however, as a kind of **estimate of the distribution of the variable in the population**; that is, we hope that aspects of the data we see in the histogram will “hold true” if we were to consider the entire population

In this sense, the default rules attempt to provide us with views of the data that **have features we can expect will exist in the population**; the different rules, then, make different assumptions about what the population data look like

Ultimately, however, these rules are derived mathematically under idealized conditions (maybe the population data have a “bell shape” -- more on that shortly) and typically for large sample sizes, making their use in practice subject to some artful interrogation

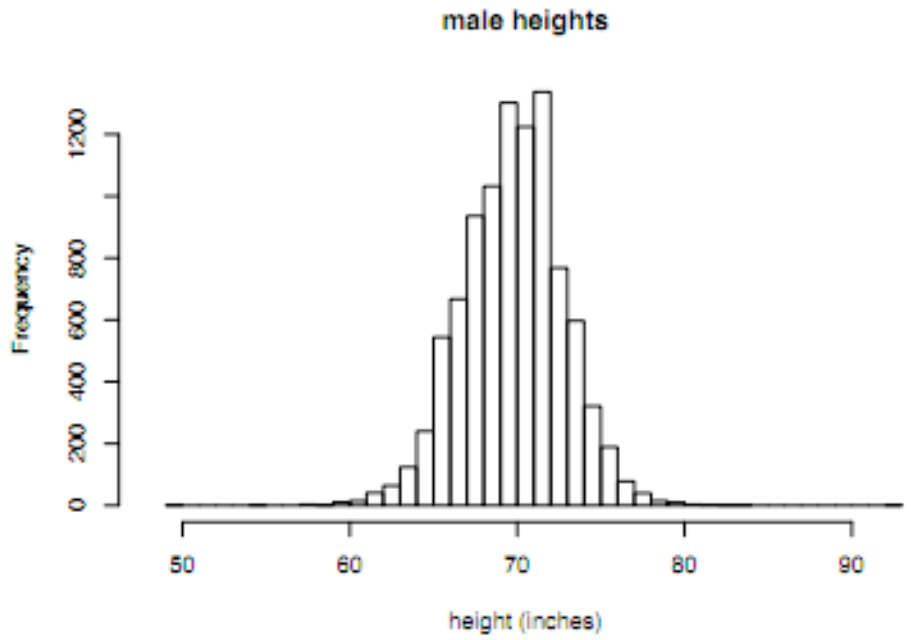
Bottom line: It’s sufficient that you understand there is a choice to be made and that reasonable defaults exist but that **you should question the defaults** if you have the time and examine several bin widths

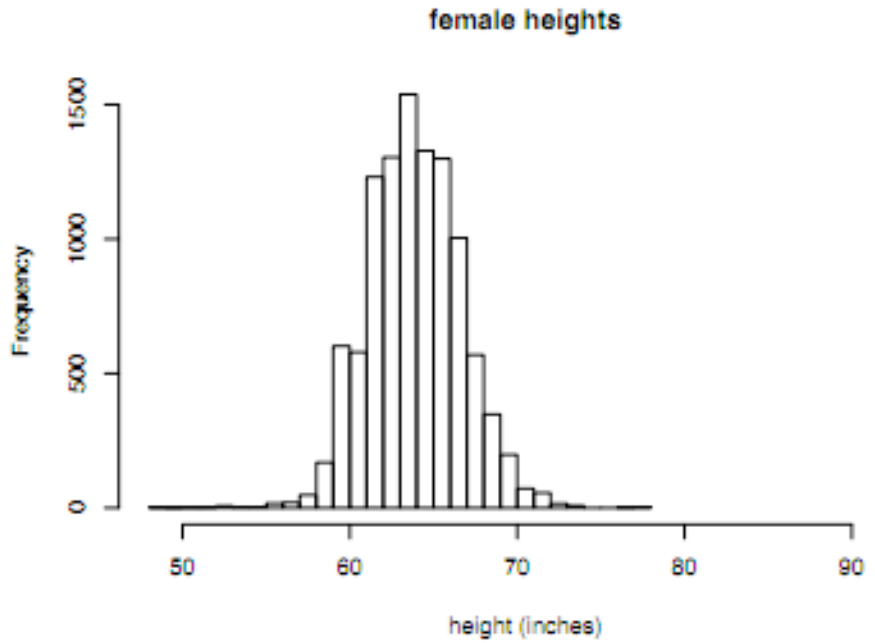
Comparing distributions (I)

We can use these displays to compare distributions

At the left we have separate histograms of the heights of males and females in the sample

What do you see? What terms would you use to describe their shapes?

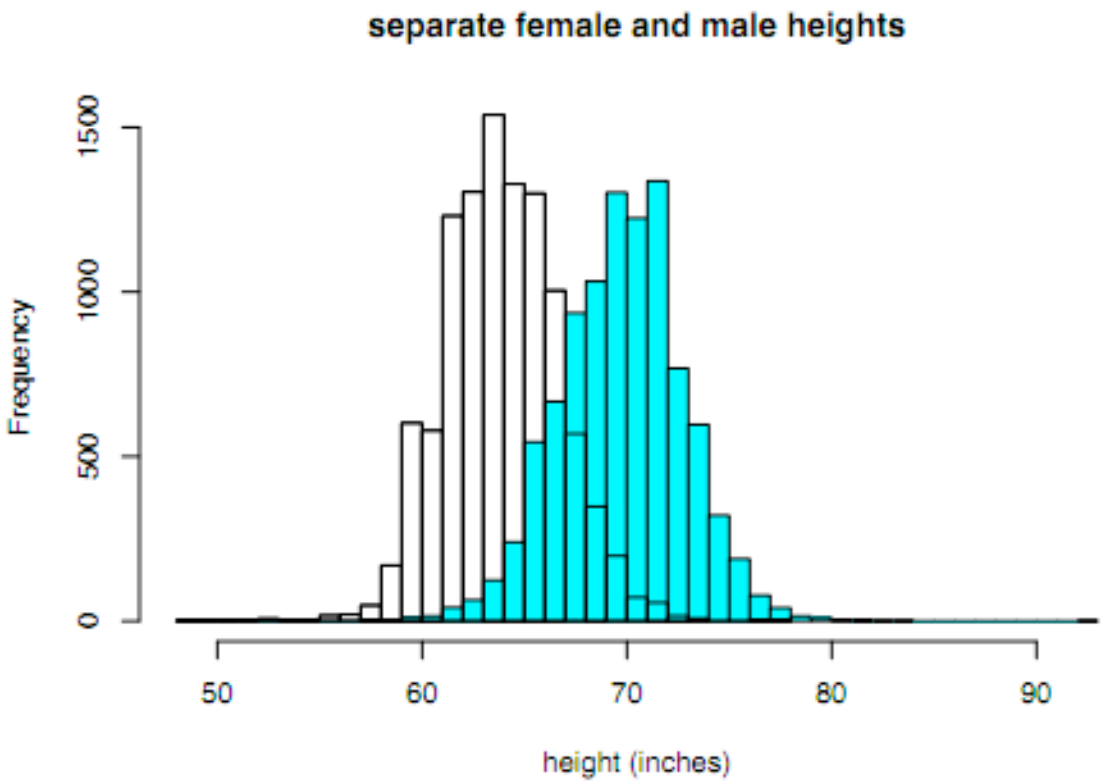




Comparing distributions (I)

A more effective strategy would be to simply overlay one histogram over the other, perhaps adding a snappy color

At this point it should be clear how helpful it is to have a good rule of thumb for picking the number of bins



An aside about skew

Body measurements are very, well, 19th century; it was all the rage at the time, fueled in part by an obsession with criminals, with identifying specific criminals or characterizing criminal "classes"

Lots of body measurements end up having "bell shaped" distributions (like people's heights), or are subject to mild skew (the BRFSS weights); a lot of work went into reasoning around the bell curve, as we'll see

RELEVÉ

DU

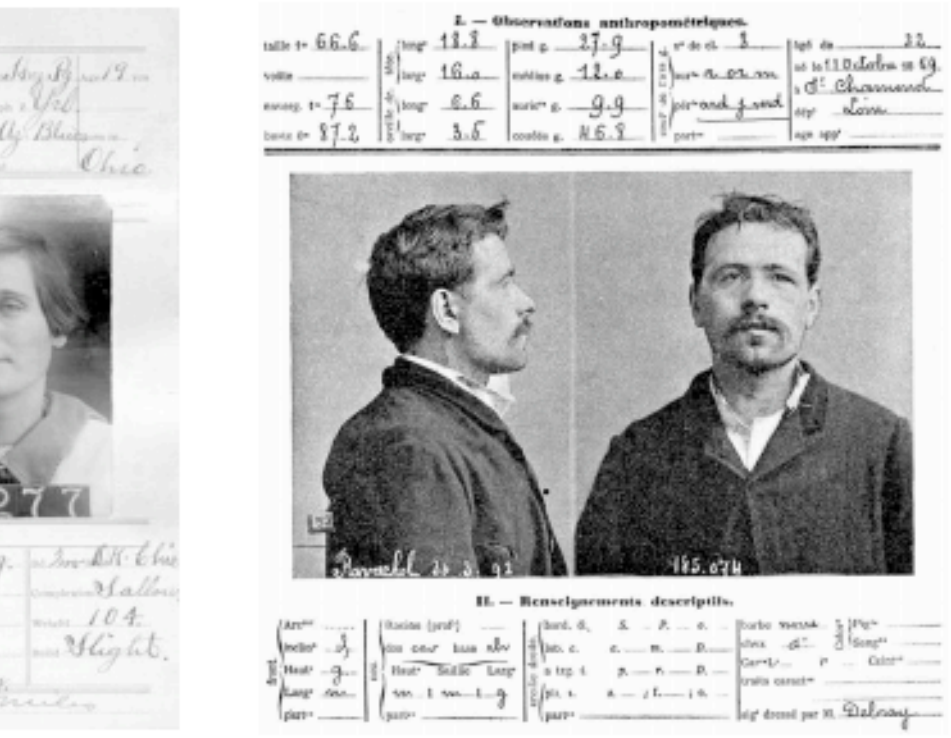
SIGNALEMENT ANTHROPOMÉTRIQUE

1. Taille. — 2. Envergure. — 3. Buste. —
4. Longueur de la tête. — 5. Largeur de la tête. — 6. Oreille droite. —
7. Pied gauche. — 8. Médius gauche. — 9. Coudée gauche.


(Réduction photographique 1/7.)
165.670


DESCRIPTIVE


DESCRIPTIVE
43277


I. — Observations anthropométriques.
II. — Renseignements descriptifs.
165.674

An aside about skew

In the late 1990s, interest was not on bells, but on skew, on extreme skew; so-called heavy tailed distributions and power laws were cropping up everywhere

As an example, let's consider measurements that are personal but not related to the body; say, the number of friends you have on Facebook or the number of visits you get to your web site or the number of bytes you download each day

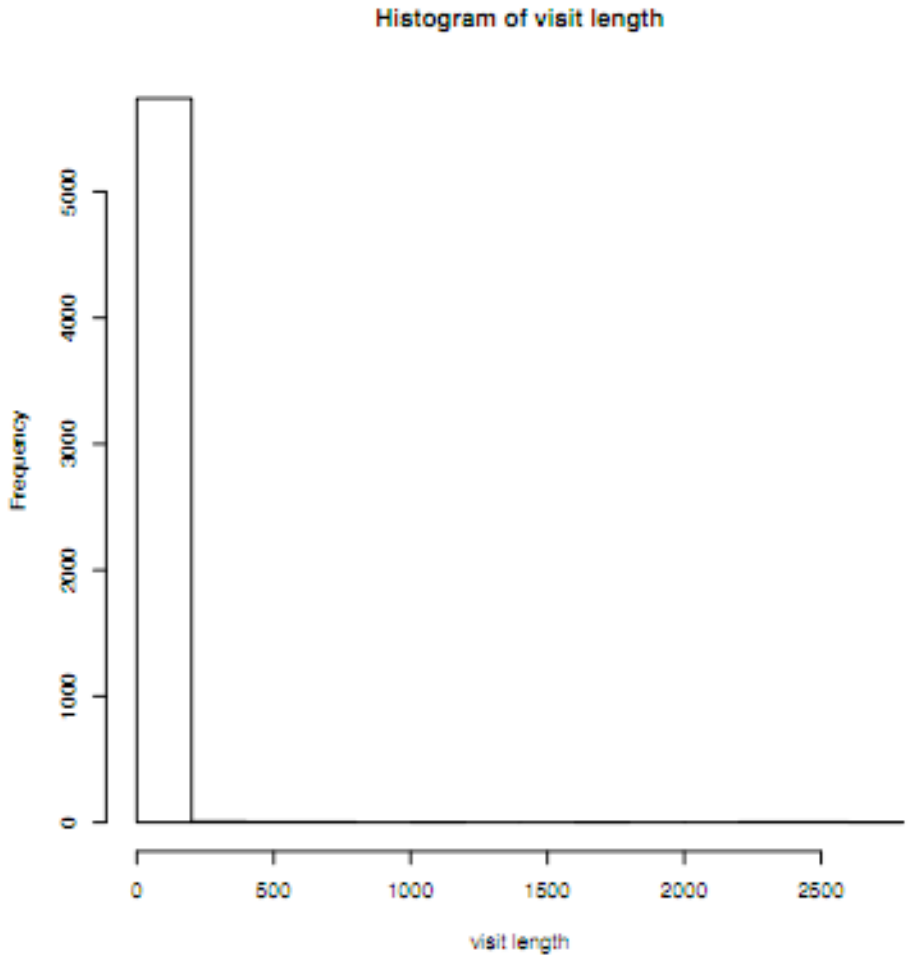
Visits to web sites

Most web sites keep records of who visits and how often; the number of unique pages viewed during a visit is called **the visit length**

For the week between 9/29/10 and 10/02/10 we collected the visit lengths associated with all 5,761 visits to `www.stat.ucla.edu`

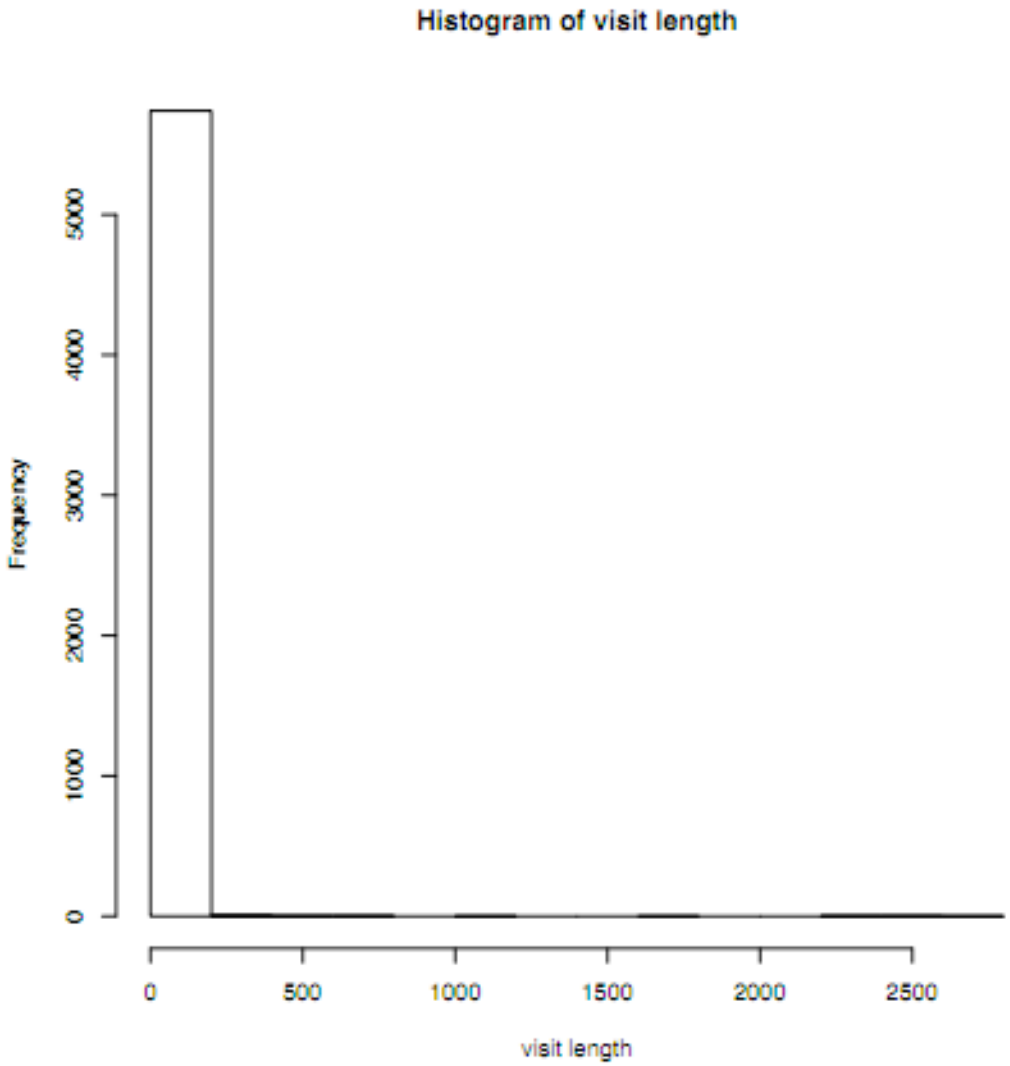
The mean visit length is 11, the median is just 3 (um, I think I just got ahead of myself there)

What does the histogram tell you?



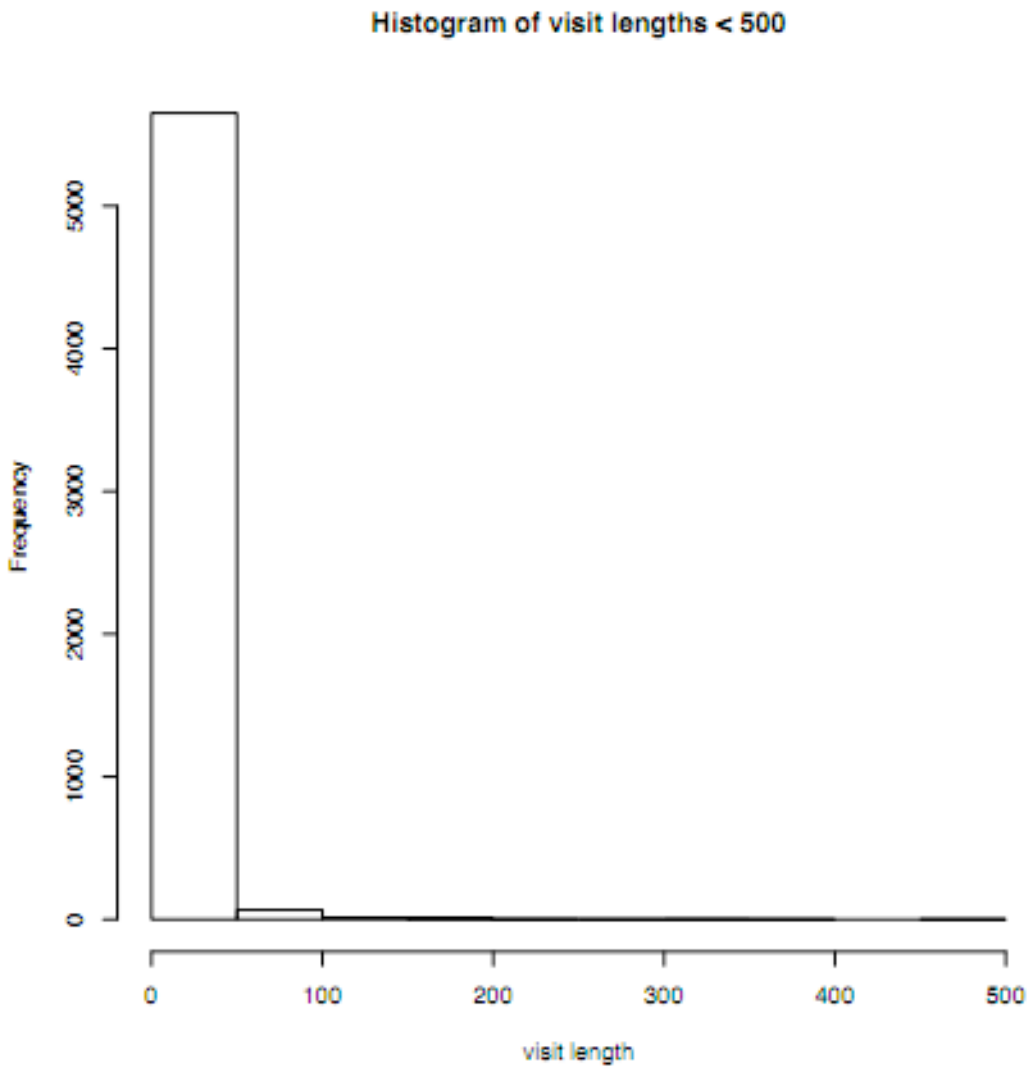
Visits to web sites

The data are badly skewed; the minimum is 1, the maximum is 2,720! We can try to restrict the range of the plot a bit...



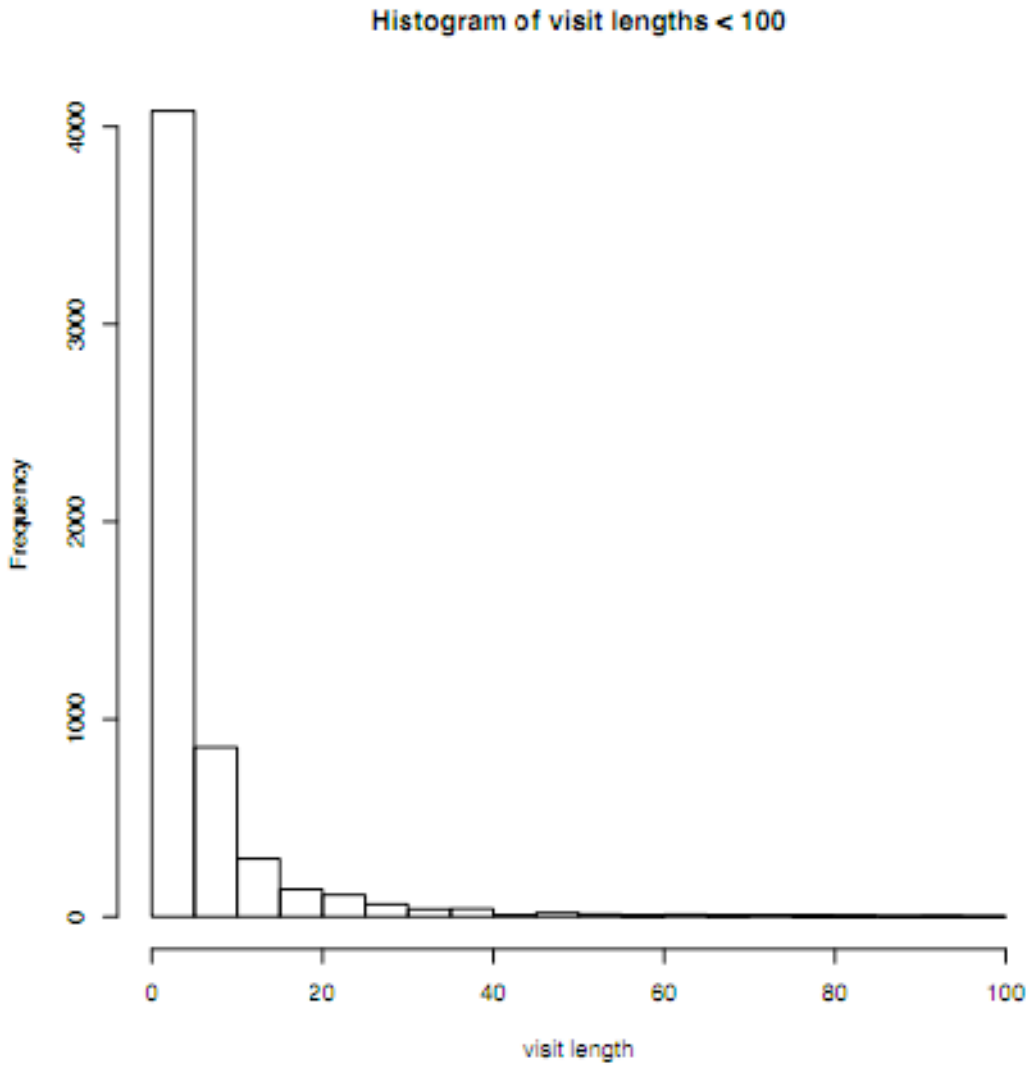
Visits to web sites

The data are badly skewed; the minimum is 1, the maximum is 2,720! We can try to restrict the range of the plot a bit...



Visits to web sites

... and then increase the number of bins; this skew is really dramatic and it will be hard to see a lot in the data this way



An aside about skew

In the late 1990s, John Doyle at CalTech and Jean Carson at UCSB launched a study of complexity and robustness

They provided a theory to explain the heavy tail distributions seen in the sizes of forest fires, in the sizes of files on the web, and a host of other systems with "highly structured, nongeneric, self-dissimilar internal configurations" that are robust yet fragile

While this topic is certainly beyond the reach of this brief introduction, suffice it to say that the world is not all about bell-shaped distributions and that in many cases we are faced with something much more extreme

In these cases, we often prefer to introduce a transformation of the data; typically we would apply a square root or a logarithm, why?

Complexity and robustness

J. M. Carlson*,† and John Doyle‡

+ Author Affiliations

Abstract

Highly optimized tolerance (HOT) was recently framework to study fundamental aspects of cor primarily by systems from biology and engineerin structured, nongeneric, self-dissimilar internal confi fragile external behavior. HOT claims these are the complexity and not accidents of evolution or artifice are inevitably intertwined and mutually reinforcing. I our paper contrasts HOT with alternative perspective real-world examples and also model systems, pa organized criticality.

A vision shared by most researchers in complex syst perhaps even universal, features capture fundamenta manner that transcends specific domains. It is in id sharp differences arise. In disciplines such as biol economics, and ecology, individual complex system of study, but there often appears to be little con models, abstractions, and methods. Highly optimiz one recent attempt, in a long history of efforts, to d for studying complexity. The HOT view is motivated and engineering. Theoretically, it builds on mathen control, communications, and computing. In this pap examples but avoid theories and mathematics th

Comparing distributions (II)

While histograms are powerful, sometimes we want to make very rough comparisons between more than two distributions

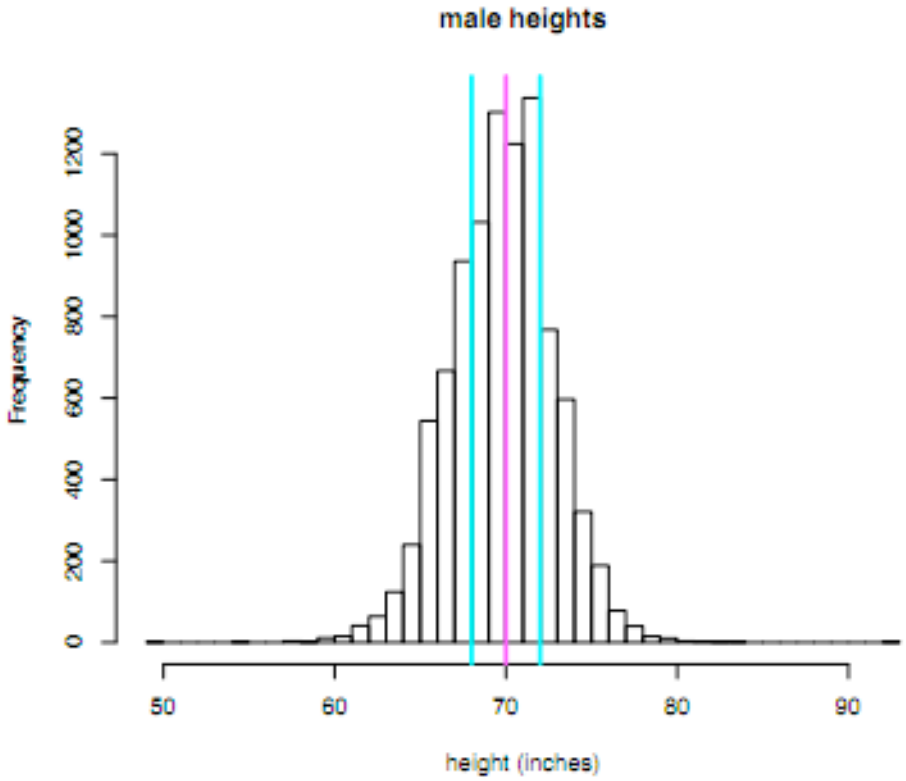
To do this, we will introduce a kind of cartoon representation of a frequency distribution for continuous data

The median and quartiles

To craft this cartoon, we begin by dividing the data in half; that is, we identify a center point

The simplest way to do this is to choose the point for which half of the data lie to the left and half to the right; this is called **the median**

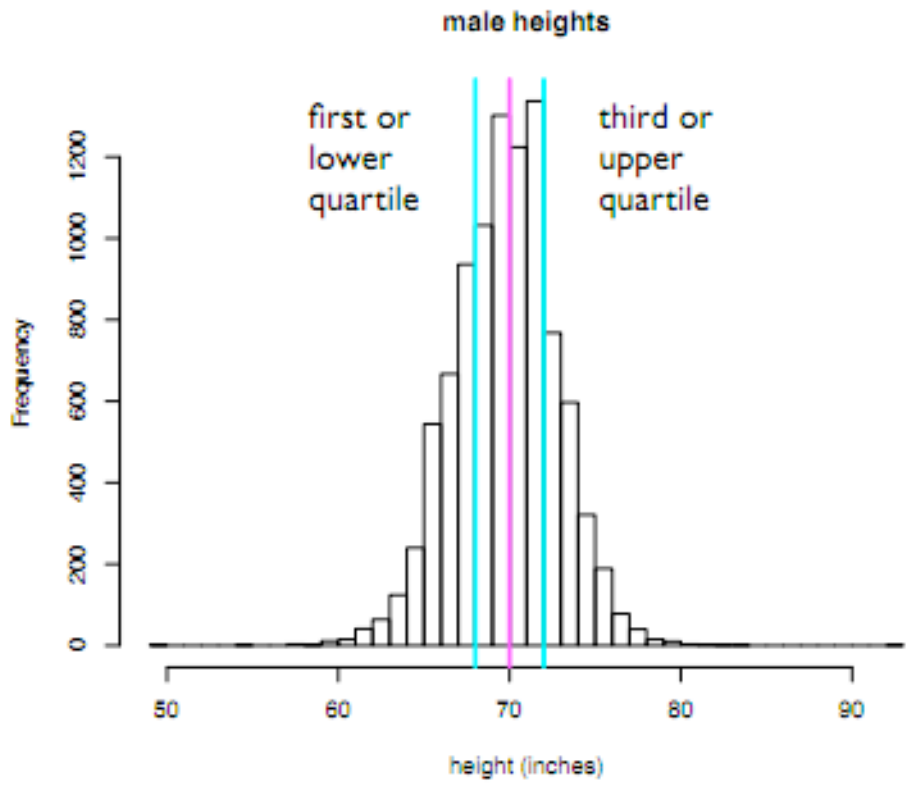
We then take the two halves and in turn divide them in half; in effect, we have split the data into four pieces



The median and quartiles

We often refer to the points marked by the cyan lines as the upper and lower quartiles

They are also called the first and third quartiles; as you might expect, the median is also known as the second quartile



Interquartile range

By design, the interval marked by the upper and lower quartiles contain half of the data; it is known as **the interquartile range**

While the median describes the center of the data, the interquartile range gives us a sense of how **spread out the data are**

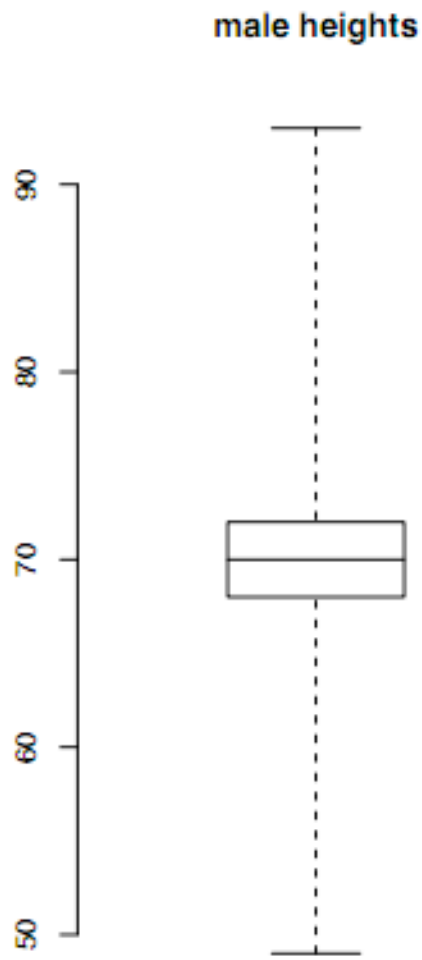
What other measures of spread might we consider?

Boxplots

The boxplot is a graphical representation of a frequency distribution; it is based on five numbers, the minimum data value, the maximum, and the three quartiles

In its standard form*, the whiskers mark the minimum and maximum values; the box is defined by the interquartile range and the median is the horizontal bar in the middle of the box

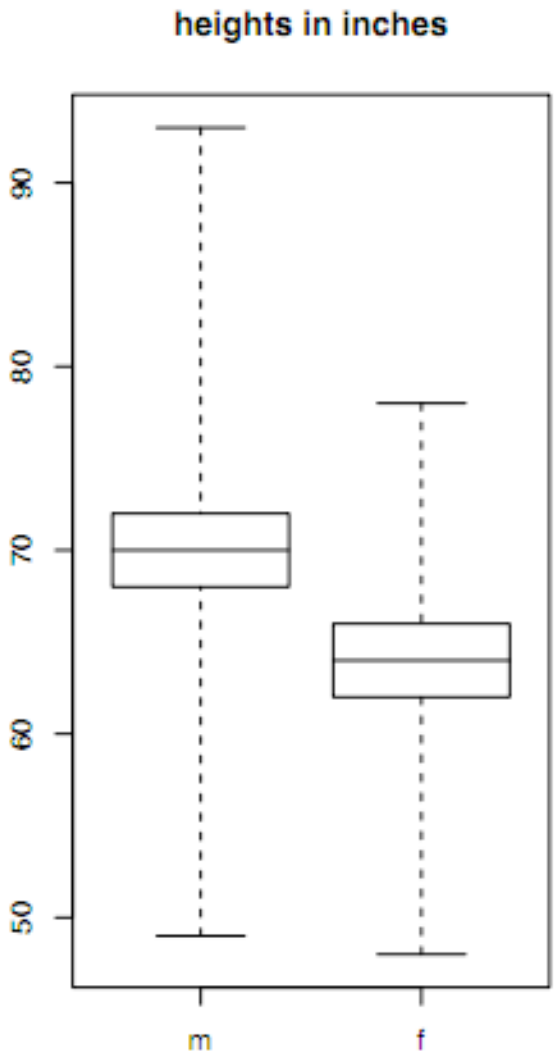
Oh, we should mention that boxplots were developed by John Tukey who we met in the first lecture

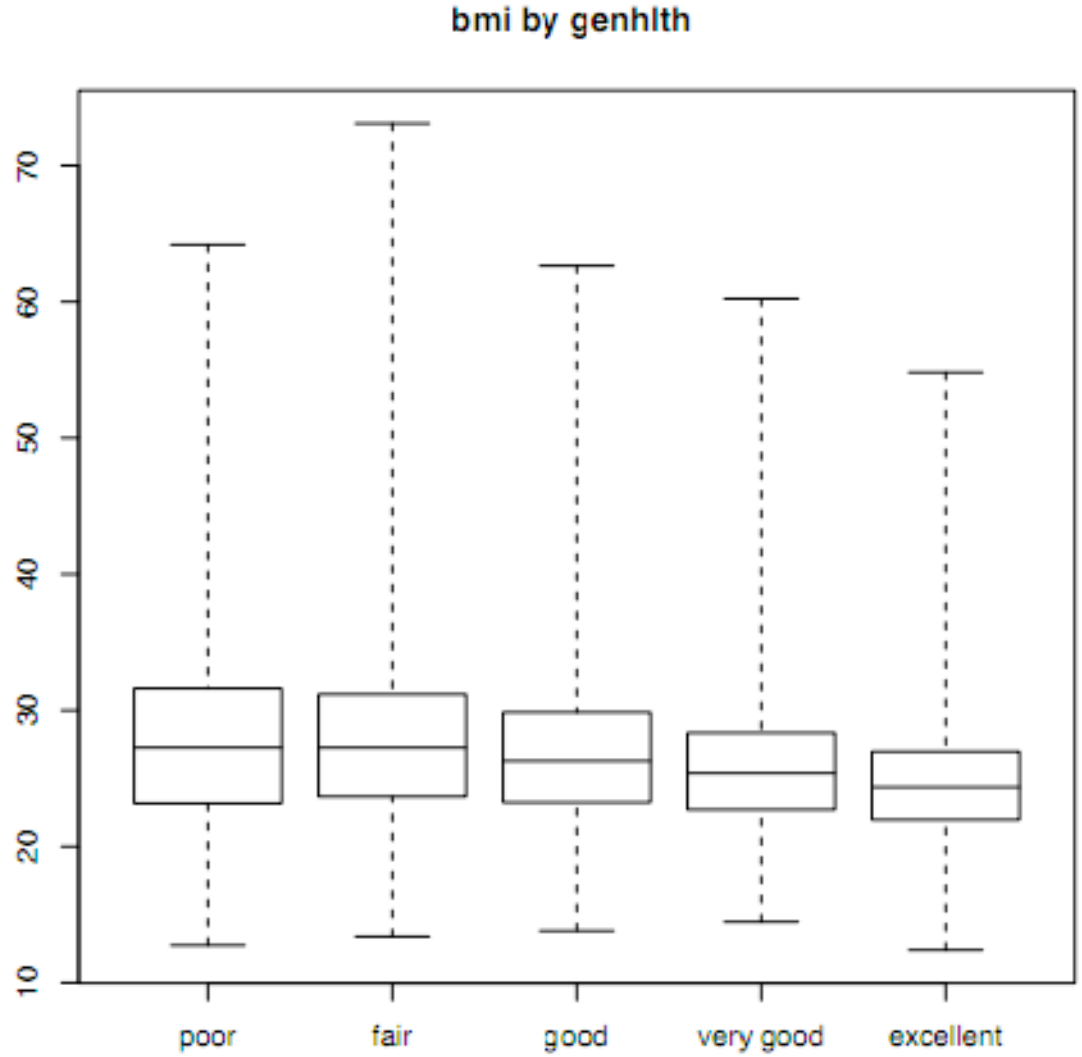


Boxplots

While they are fairly cartoonish, boxplots let us compare a number of distributions at one time

Here we have male and female heights...




bmi by genhlth
70
60
50
40
30
20
10
poor
fair
good
very good
excellent

Boxplots

R implements a modified version of the boxplot, one that incorporates the notion of **outliers**

Outliers are points that stand out from the rest of the data in some way; we typically identify such points on the basis of our prior expectations about how data should behave

Modified boxplots

The boxplots we have seen so far are direct graphical representations of the so-called five number summary

Given a sample of observations on a quantitative variable, the five number summary consists of **the minimum, the lower quartile, the median, the upper quartile and the maximum**

The default boxplot provided by most (if not all) modern statistical software packages is a little more complex; it attempts to highlight values that are **“too extreme”**

Tukey's motivation

*A cautious data analyst often has reason for concern over the distortions that aberrant observations can cause... It is informative, and may be important, to examine samples... for the presence of "outliers" or "exotic values" because their unexpected behavior may indicate failure of a model or point to an unanticipated phenomenon.**

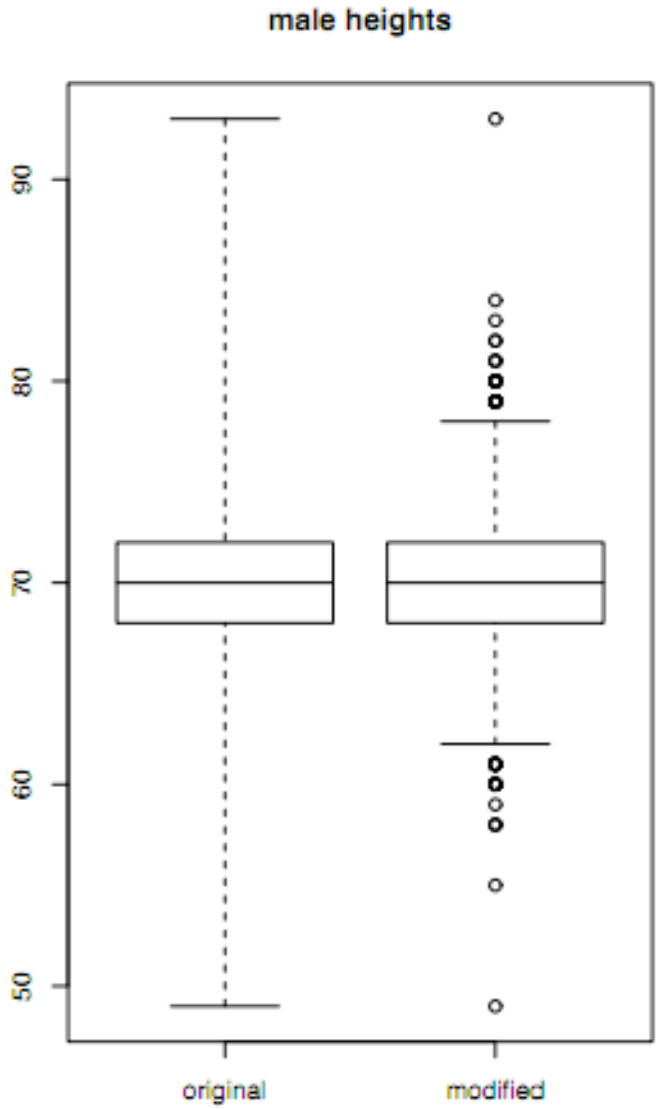
* David C. Hoaglin, Boris Iglewicz, and John W. Tukey (1986), Performance of some resistant rules for outlier labeling, *Journal of the American Statistical Association*, Vol. 81, pp. 991-999.

Modified boxplots

Here we have two boxplots for the heights of males in the CDC study from last lecture; the plot on the left is the original boxplot we created last time, while the one on the right is its "modified" cousin

In this case, the new plot brings the whiskers closer to the middle of the display; by peeling back the whiskers, we expose some of the actual data beyond the "fences"

What does this give us?



Tukey's simple rule

Recall that the distance between the upper and lower quartiles is known as the **interquartile range** or IQR

Define **lower fence** to be $Q_1 - 1.5\,\text{IQR}$; we say that any point below the lower fence is a possible outlier, requiring some investigation

Define the **upper fence** to be $Q_3 + 1.5\,\text{IQR}$; we say that any point above the upper fence is also a possible outlier

Tukey's simple rule

We then highlight potential outliers in our boxplots by adjusting the placement of the upper and lower whiskers

If there are no data points below the lower fence, we leave the lower whisker at the minimum data point; if there are, we place the whisker at the smallest point above the lower fence

If there are no data points above the upper fence, we leave the upper whisker at the maximum data point; if there are, we place the whisker at the largest point below the upper fence

Example

For males, the five number summary of the variable weight is

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
78.0	165.0	185.0	189.3	210.0	500.0

Therefore, the IQR is (210-165) = 45 and our fences are computed to be

lower: 165 - 1.5*45 = 97.5

upper: 210 + 1.5*45 = 277.5

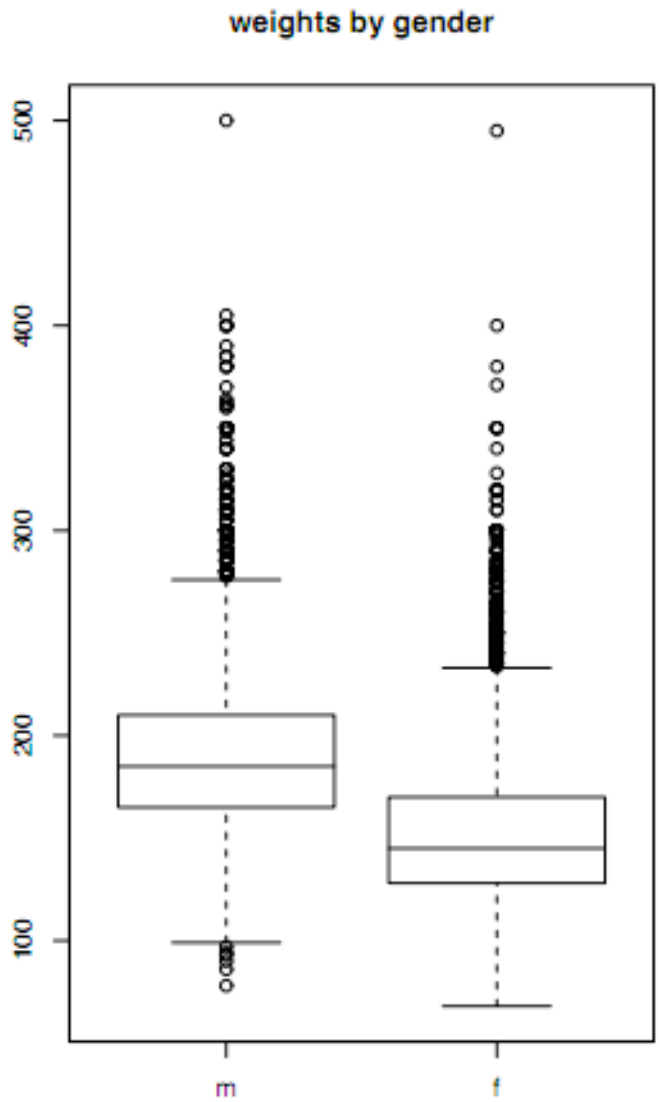
Using this, we identify 6 possible outliers that are extreme and small, and 262 that are extreme and large

Modified boxplots

What do we think of this alteration? What new information does it provide? How might this be useful for us when examining data?

The choice to highlight points falling above or below 1.5 times the IQR, is just that, a choice; why not 1.0 or 3.0 times the IQR, how would those choices change things?

How do you think the value of 1.5 was settled on? What principles were used to guide the choice? Maybe an easier question is, how do we want this



Performance of Some Resistant Rules for Outlier Labeling

DAVID C. HOAGLIN, BORIS IGLEWICZ, and JOHN W. TUKEY*

The techniques of exploratory data analysis include a resistant rule for identifying possible outliers in univariate data. Using the lower and upper fourths, F_L and F_U (approximate quartiles), it labels as "outside" any observations below $F_L - 1.5(F_U - F_L)$ or above $F_U + 1.5(F_U - F_L)$. For example, in the ordered sample $-5, -2, 0, 1, 8$, $F_L = -2$ and $F_U = 1$, so any observation below -6.5 or above 5.5 is outside. Thus the rule labels 8 as outside. Some related rules also use cutoffs of the form $F_L - k(F_U - F_L)$ and $F_U + k(F_U - F_L)$. This approach avoids the need to specify the number of possible outliers in advance; as long as they are not too numerous, any outliers do not affect the location of the cutoffs.

To describe the performance of these rules, we define the some-outside rate per sample as the probability that a sample will contain one or more outside observations. Its complement is the all-inside rate per sample. We also define the outside rate per observation as the average fraction of outside observations. For Gaussian data the population all-inside rate per sample (0) and the population outside rate per observation (.7%) substantially understate the corresponding small-sample values. Simulation studies using Gaussian samples with n between 5 and 300 yield detailed information on the resistant rules. The main resistant rule ($k = 1.5$) has an all-inside rate per sample between 67% and 86% for $5 \le n \le 20$, and corresponding estimates of its outside rate per observation range from 8.6% to 1.7%.

Both characteristics vary with n in ways that lead to good empirical approximations. Because of the way in which the fourths are defined, the sample sizes separate into four classes, according to whether dividing n by 4 leaves a remainder of 0, 1, 2, or 3. Within these four classes the all-inside rate per sample shows a roughly linear decrease with n over the range $9 \le n \le 50$, and the outside rate per observation decreases linearly in $1/n$ for $n \ge 9$.

A more theoretical approximation for the all-inside rate per sample works with the order statistics $X_{(1)} \le \cdots \le X_{(n)}$. In this notation the fourths are $X_{(f)}$ and $X_{(n+1-f)}$ with $f = \text{i}[(n + 3)/2]$, where $[\cdot]$ is the greatest-integer function. A sample has no observations outside whenever $\{X_{(f)} - X_{(1)}\}/\{X_{(n+1-f)} - X_{(f)}\} \le k$ and $\{X_{(n)} - X_{(n+1-f)}\}/\{X_{(n+1-f)} - X_{(f)}\} \le k$. We first approximate the numerators and denominator in these ratios by constant multiples of chi-squared random variables with the same mean and variance. We then approximate the logarithm of each ratio by a Gaussian random variable, and we calculate the correlation between these variables from the fact that the ratios have the same denominator. Finally, a bivariate Gaussian probability calculation yields the approximate all-inside rate per sample. The error of the result relative to the simulation estimate is typically from 1% to 2% for $5 \le n \le 50$.

To provide an indication of how the two rates behave in alternative "null" situations, the simulation studies included samples from five heavier-tailed members of the family of h-distributions. For a given sample size, the all-inside rate per sample decreases as the tails become heavier, and the outside rate per observation increases.

KEY WORDS: Bivariate normal distribution; Chi-squared distribution; Exploratory data analysis; h-distributions; Masking.

1. INTRODUCTION

A cautious data analyst often has reason for concern over the distortions that aberrant observations can cause. By using summaries that change only slightly in response to an arbitrary change in any small part of the data, robust and resistant methods have made it possible to minimize the effects of such unusual data. It is still informative, however, and may be important, to examine samples and residuals for the presence of "outliers" or "exotic values" because their unexpected behavior may indicate failure of a model or point to an unanticipated phenomenon.

Research on methods of testing for outliers has produced an extensive literature, discussed in books by Barnett and Lewis (1978, 1984) and Hawkins (1980) and in articles by Barnett (1983) and Beckman and Cook (1983). A number of the proposed procedures have difficulty when a sample may contain multiple outliers. The problems include *masking*, in which the presence of other outliers makes each outlier difficult to detect, and *swamping*, in which the procedure tends to declare too many outliers when the null hypothesis of no outliers is rejected. From the point of view of robust/resistant data analysis, masking occurs because the procedure has a too-low breakdown point. [Donoho and Huber (1983) discussed the breakdown point in some detail. They defined it as "roughly, the smallest amount of contamination that may cause an estimator to take on arbitrarily large aberrant values" (p. 157).] Thus, by having higher breakdown points, resistant measures of a sample's location and spread should make it possible to avoid masking in most situations.

As one informal step in this direction, exploratory data analysis (Tukey 1977a) includes a resistant rule of thumb for identifying observations that are extreme and hence are potential outliers. The breakdown point of this rule is roughly 25%. Its practical advantages include simplicity, ability to identify multiple outliers, and routine use in such displays as boxplots. Thus it is valuable to study the probability behavior of the rule in "null" and "alternative" situations.

Small to moderately large samples from the Gaussian distribution constitute the customary null situation. We use simulation to measure the rule's performance in terms of three criteria: (a) the all-inside rate per sample, (b) the probability that (in small samples) as many as three of the observations are "outside," and (c) the outside rate per

* David C. Hoaglin is Research Associate, Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138. Boris Iglewicz is Professor, Department of Statistics, Temple University, Philadelphia, PA 19122. John W. Tukey is Senior Research Statistician and Donner Professor of Science Emeritus, Fine Hall, Princeton University, Princeton, NJ 08544. The authors are grateful to Gerry Feit, Jorge Martinez, and Gaston Mendoza for their assistance in carrying out the simulation work, to A. R. DiDonato for supplying the program used to calculate bivariate normal probabilities, and to Michael Dolker, John D. Emerson, Peter J. Kempthorne, Frederick Mosteller, James L. Rosenberger, Roy E. Welsch, Cleo Youtz, an associate editor, and two referees for helpful comments. This work was supported in part by U.S. Army Research Office Contract DAAG29-82-K-0085 with Harvard University, by National Science Foundation Grant SOC75-15702 to Harvard University, by National Cancer Institute Grant CA-31247 to Harvard University, and by U.S. Army Research Office Contract DAAG29-82-K-0178 with Princeton University. An earlier version of this article appeared in the *1980 Proceedings of the Statistical Computing Section.*

But, it's just a graphical tool...

We could examine how far out to draw the fences by **trying out different choices in situations where we know the answer** (well behaved data with no anomalous points, data with a few odd points) and see what happens

This is precisely what Tukey does in his paper; **he simulates data from a model** where there are no strange points (a bell-shaped distribution, actually, but we'll get to that later) and tries to make sure that in those "normal" cases, the plot doesn't flag too many points as possibly anomalous

Tukey and company comment that...

> *Any observations that fall below the [lower fence] or above [the upper fence] are termed outside. We would inquire into the circumstances surrounding any outside data values in an attempt to learn the reasons for their unusual behavior, and we are likely to level the corresponding points with appropriate identifiers in any graphical display.*

But keep in mind that **there is a difference between a point being "outside," or an "outlier" as is it is commonly called, and there being a real problem with the data represented by that point**; the modified box plot is meant to call your attention to data that might be strange, **it is not a rule defining what is strange or a problem**

The upshot

This is now the second time you've seen **a kind of tuning parameter** (the bin width of a histogram and the fence distance of a boxplot) that you can easily vary thanks to software

No matter whether you use the default choices (which you will typically do for boxplots, whereas for histograms, experimentation is encouraged) or not, it's important to remember that they are guides; **they are nothing more than devices that tend to make the graphics we're considering informative** in a large number of cases

So, while a graphic may indicate some points are outliers, it is up to you, the data analyst, to have a look and make a determination; at this point, all we've presented are guides that help you see things, **they are not reasoning for you!**

Extensions (I)

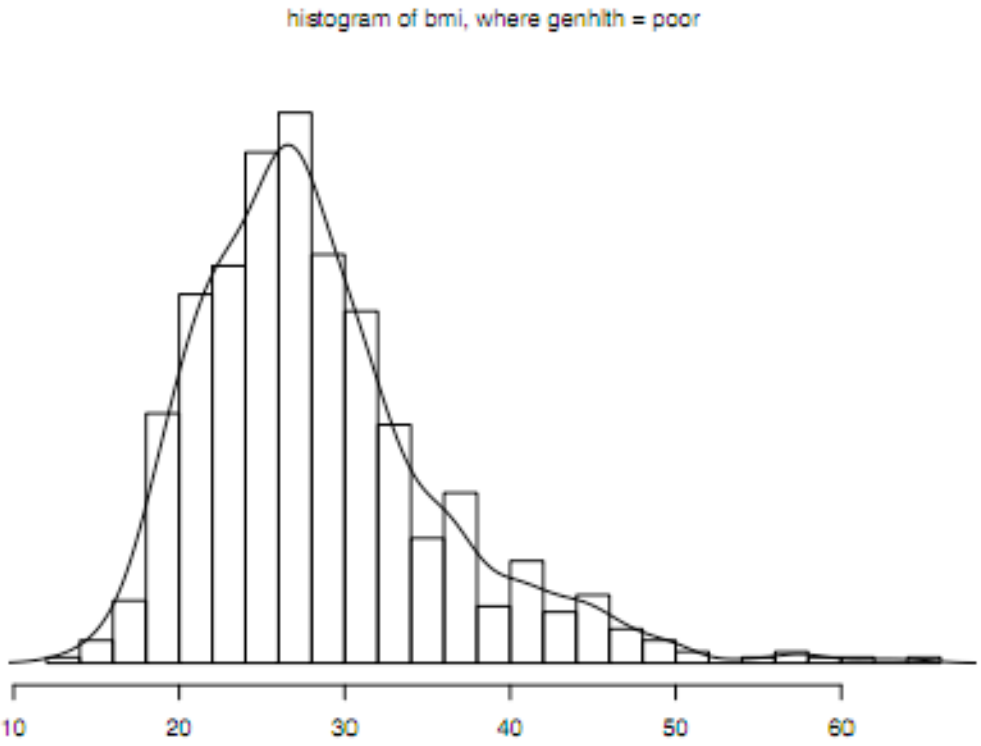
Boxplots represent a rather extreme compression of the frequency distribution; after all, only 5 numbers are being shown

They were also developed at a time when EDA might be done by hand (Tukey advocates extensive use of tracing paper, for example)

An extension of the boxplot attempts to make more of the distribution visible; it starts with a "smoothed" histogram

On the right we have a histogram of BMI for people who reported being in poor health; the curve is the smoothed histogram

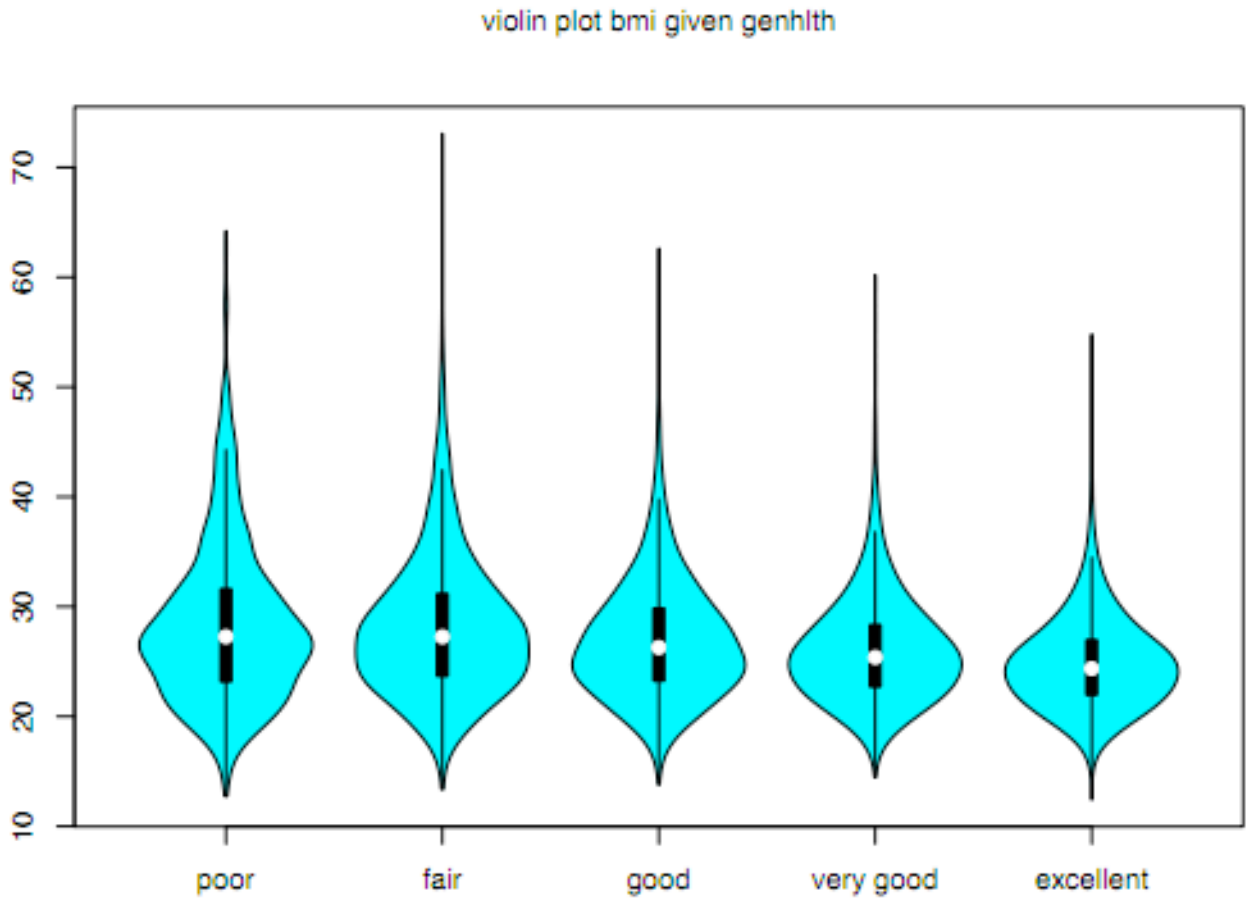
How can we use this to make the boxplot a bit more expressive?



Violin plots

The so-called violin plot might be more artistry than data analysis; but it uses the smoothed histogram tipped on its side and mirrored left and right in place of a box

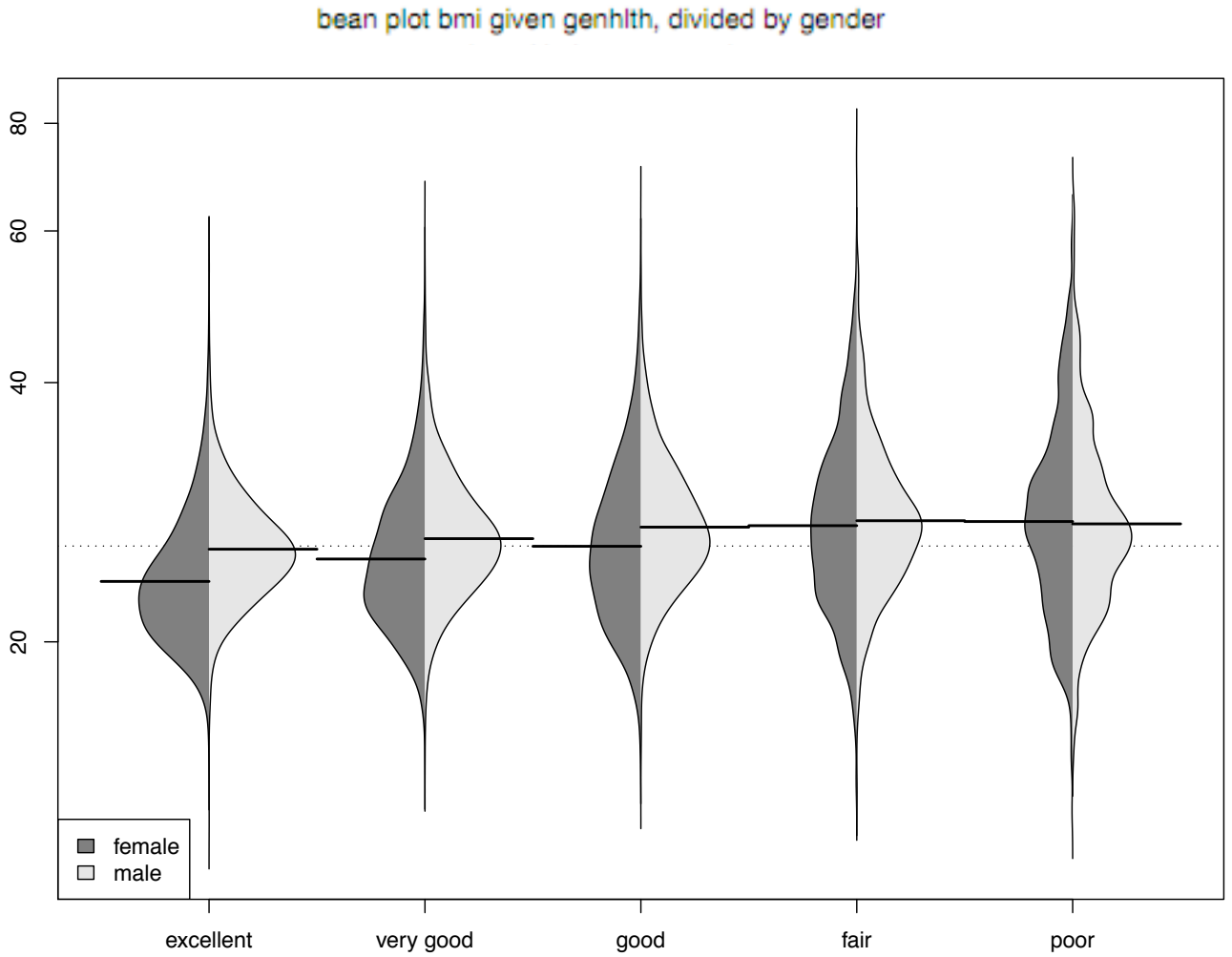
Compare this plot to the boxplot three slides back; what do you think?



Bean plots

This might be getting a bit farther out there, but with a relative of the violin plot (the so-called bean plot, named because the plots look like, um, beans) you can go farther and condition on, say, gender, producing a side-by-side plot

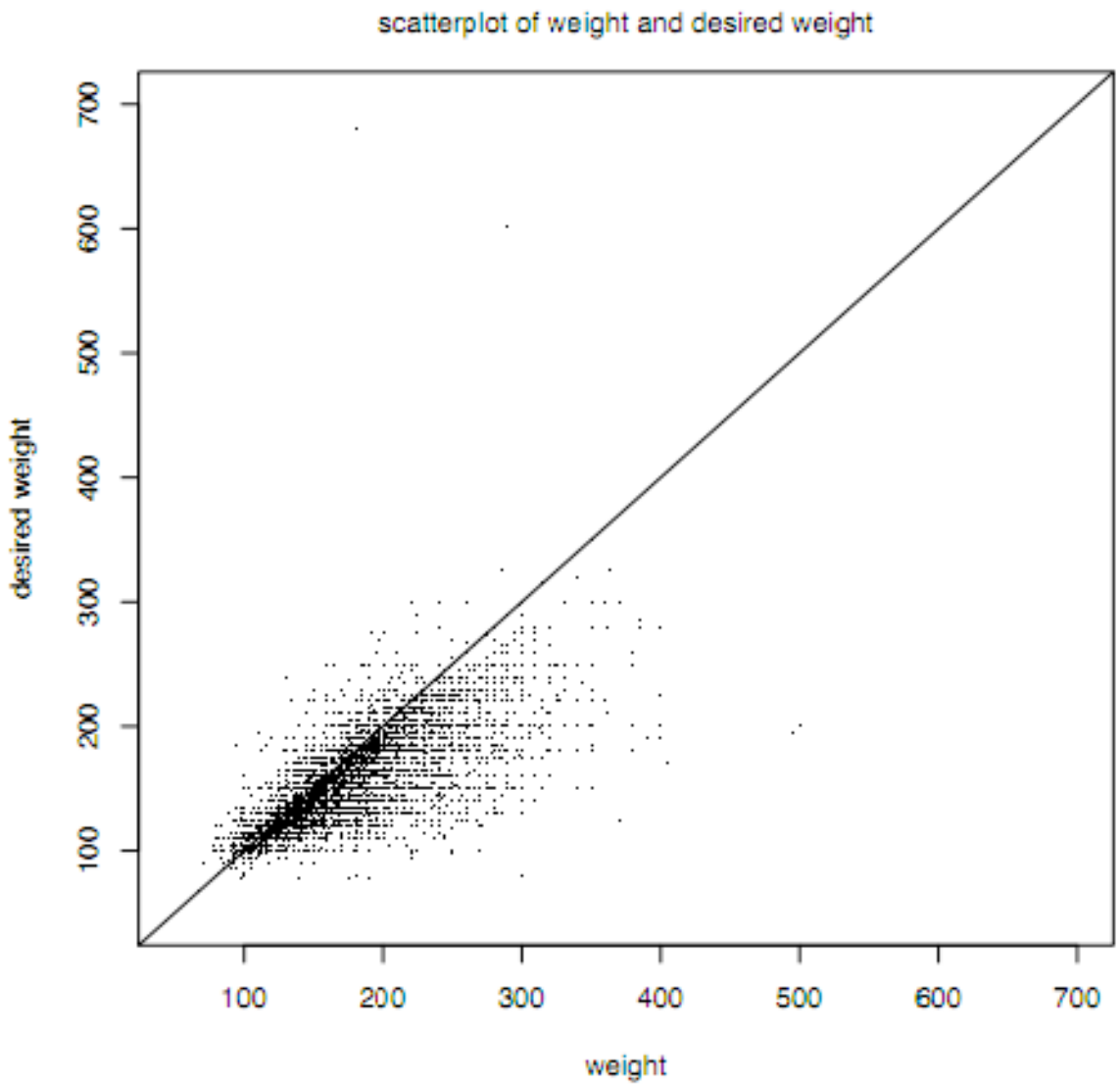
OK, now what do you think?



Extensions (II)

Now, let's consider another kind of extension to the boxplot; suppose we have two variables that we would like to describe

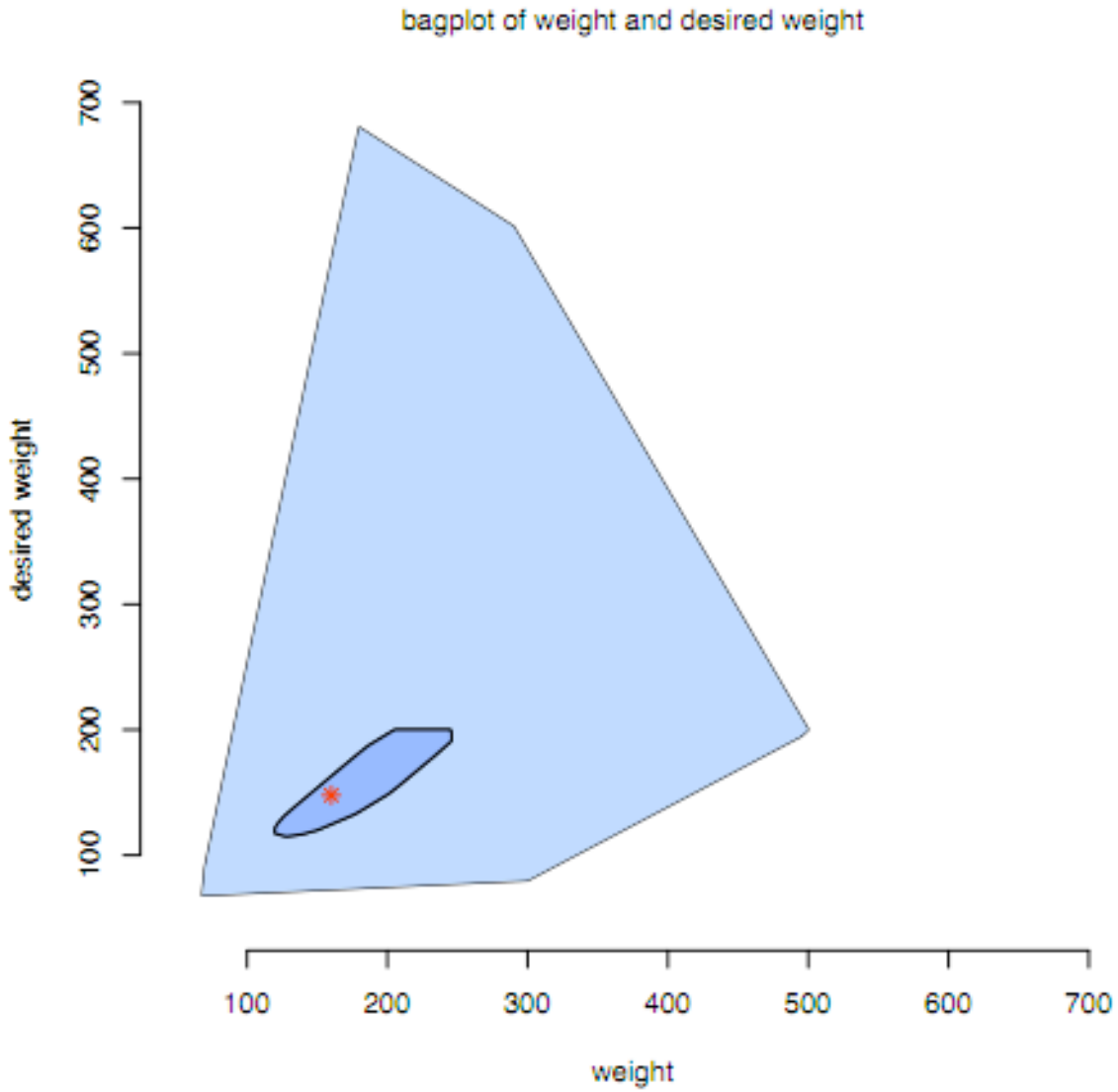
Take, for example, `weight` and `wtdesire`; we can create a simple **scatterplot** to look at how these two variables relate to each other

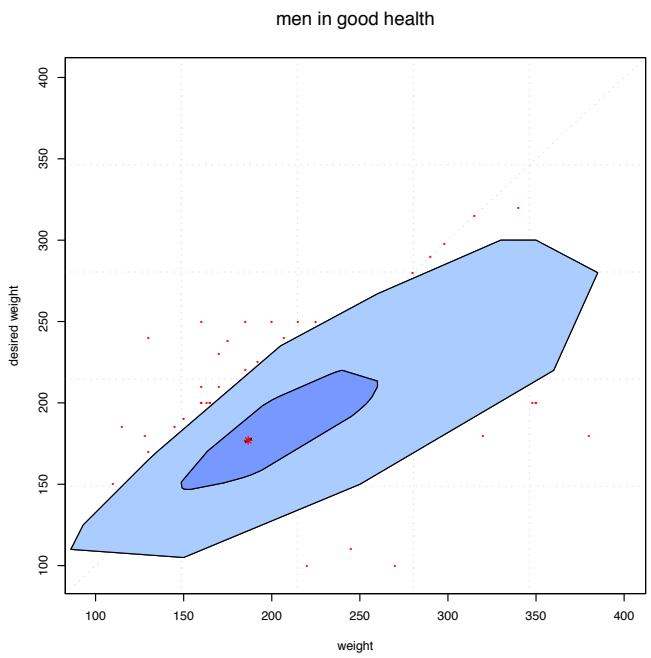

scatterplot of weight and desired weight
desired weight
weight
100
200
300
400
500
600
700

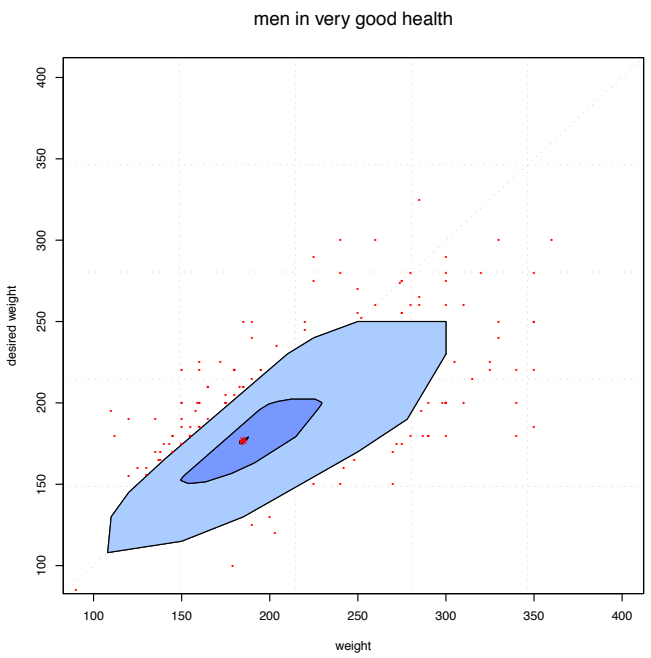
Scatterplots

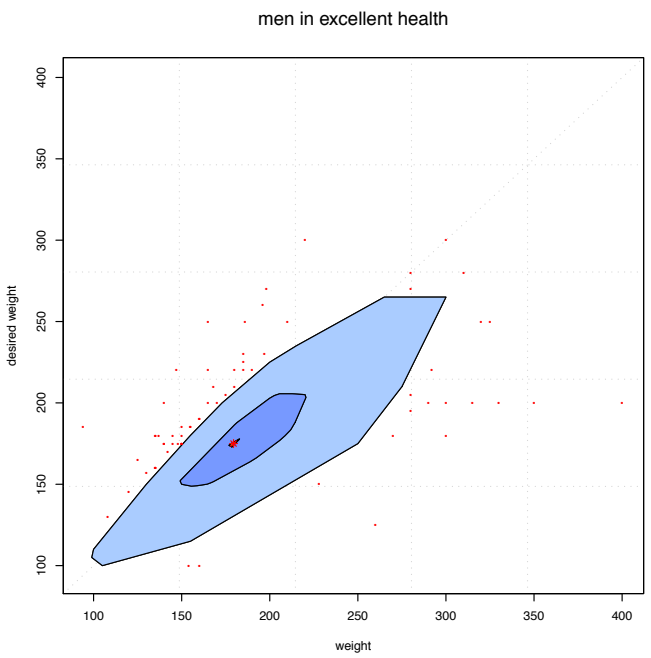
We've added a line with unit slope to the plot; Why? What do you notice? What strikes you as expected? Unexpected?

Now, suppose we want to create something like a boxplot for these data; what concepts do we have to extend?


bagplot of weight and desired weight
desired weight
100
200
300
400
500
600
700
weight
100
200
300
400
500
600
700


men in good health
desired weight
weight
100
150
200
250
300
350
400


men in very good health
desired weight
weight
100
150
200
250
300
350
400


men in excellent health
desired weight
weight
100
150
200
250
300
350
400

Bagplots

Bagplots are another Tukey innovation (along with the boxplot), but somehow they haven't caught on; why?

Can you see this being useful? Under what circumstances? How might they be interesting for our data?

Other solutions

There are, of course, other ways to address the overplotting issues we saw in our original scatterplot -- In a later lecture (or maybe lab) we'll use hexagon tiling of the graphing region to create a kind of two-dimensional histogram

But we'll save that for a later day -- For now, a few slides on other notions of center and spread...

Other measures of center and spread

So far, we have used techniques that are more about sorting than anything else; we found the "center" of a distribution by selecting a point in the middle of a sample, the median

Our notion of "spread" came from the IQR, the interquartile range; this is the central region containing 50% of the data

Over the next few slides, we illustrate two other measures, **the mean and variance**; I'd like you to read this, together with the sections in your book, but we won't talk about it much in class for another week when we start using these numerical summaries for inference

Mean and variance

A perhaps more primary notion of center for a sample of data is the **arithmetic mean or average;** given observations $x_1, x_2, \ldots, x_n$, this is just

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

In a similar spirit, **the variance** of a sample of data is the sum of squared deviations from the mean; that is, for each point x_1, the deviation is $(x_1 - \bar{x})$ and the variance s^2 is

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n\text{-}1}$$

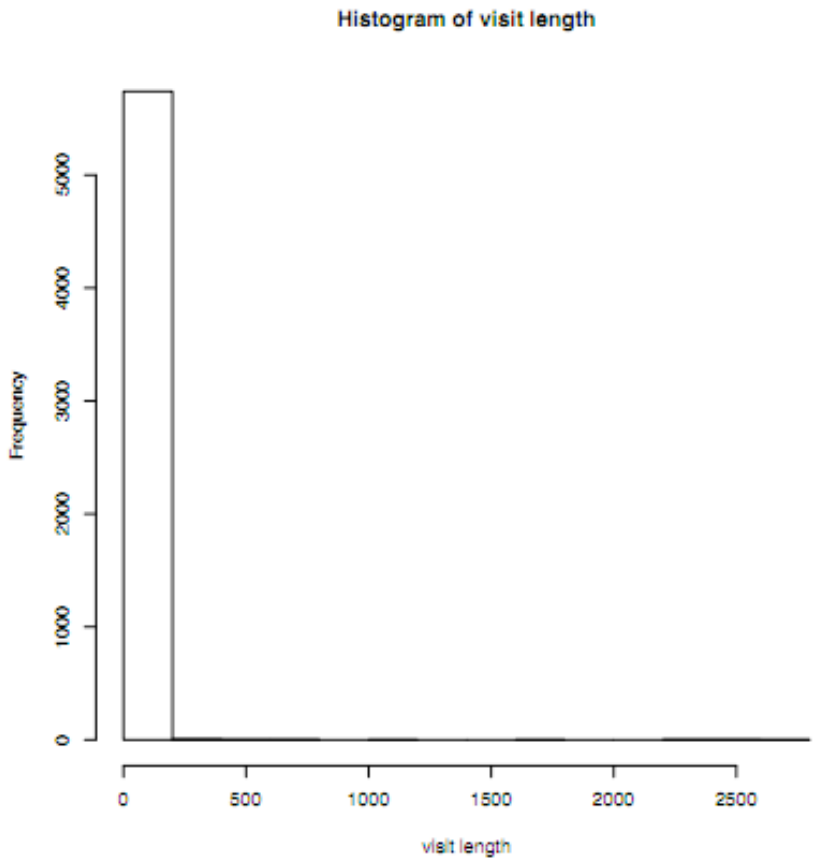
Mean and variance

While the median and the IQR were very direction notions of center and spread, the mean and standard deviation are slightly more delicate; for example, the mean is very much influenced by one or more “extreme” points (in the sense of extreme that we discussed earlier)

Why would we expect that? Is the median similarly vexed? Why or why not?

Web visits example

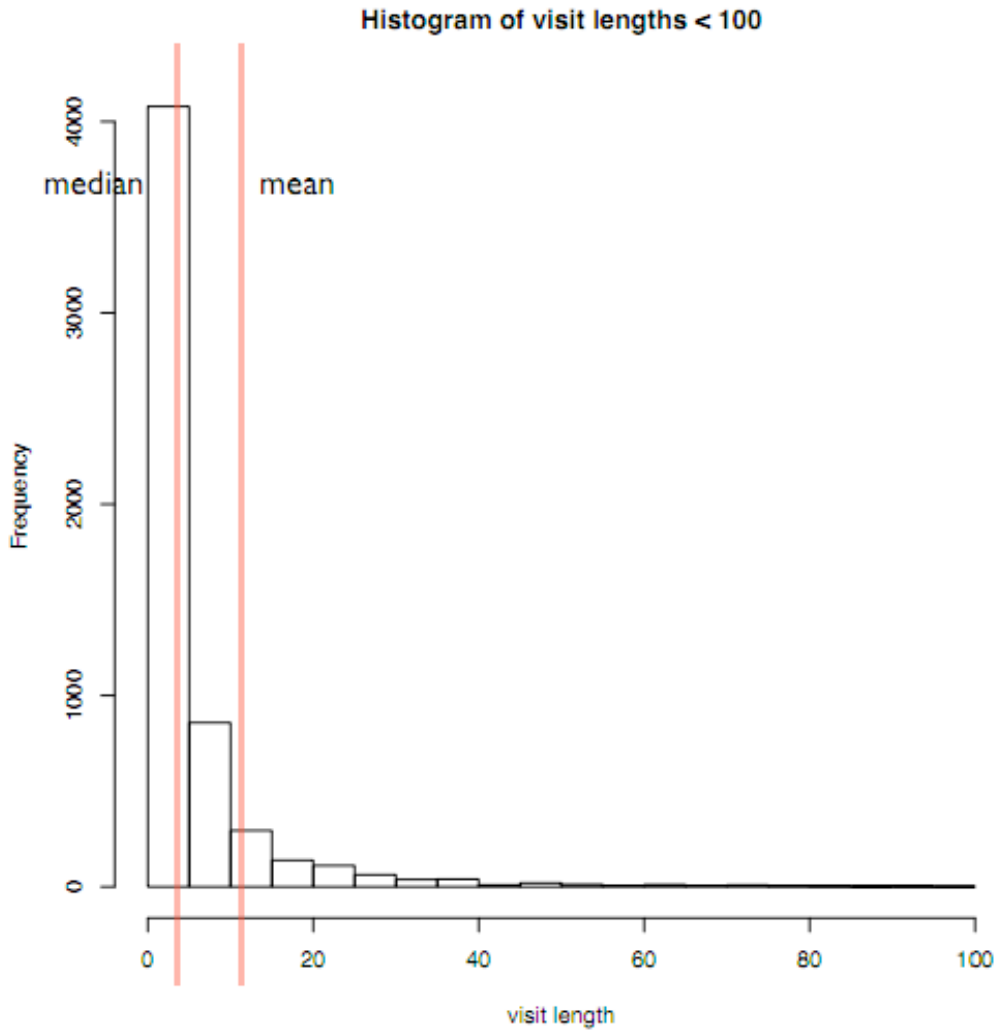
Here again is our poor histogram of web visits; remember that these data are heavily skewed to the right



Web visits example

Here we is the histogram restricted to visits of length 100

The two red lines mark the median and the mean; what do you notice?

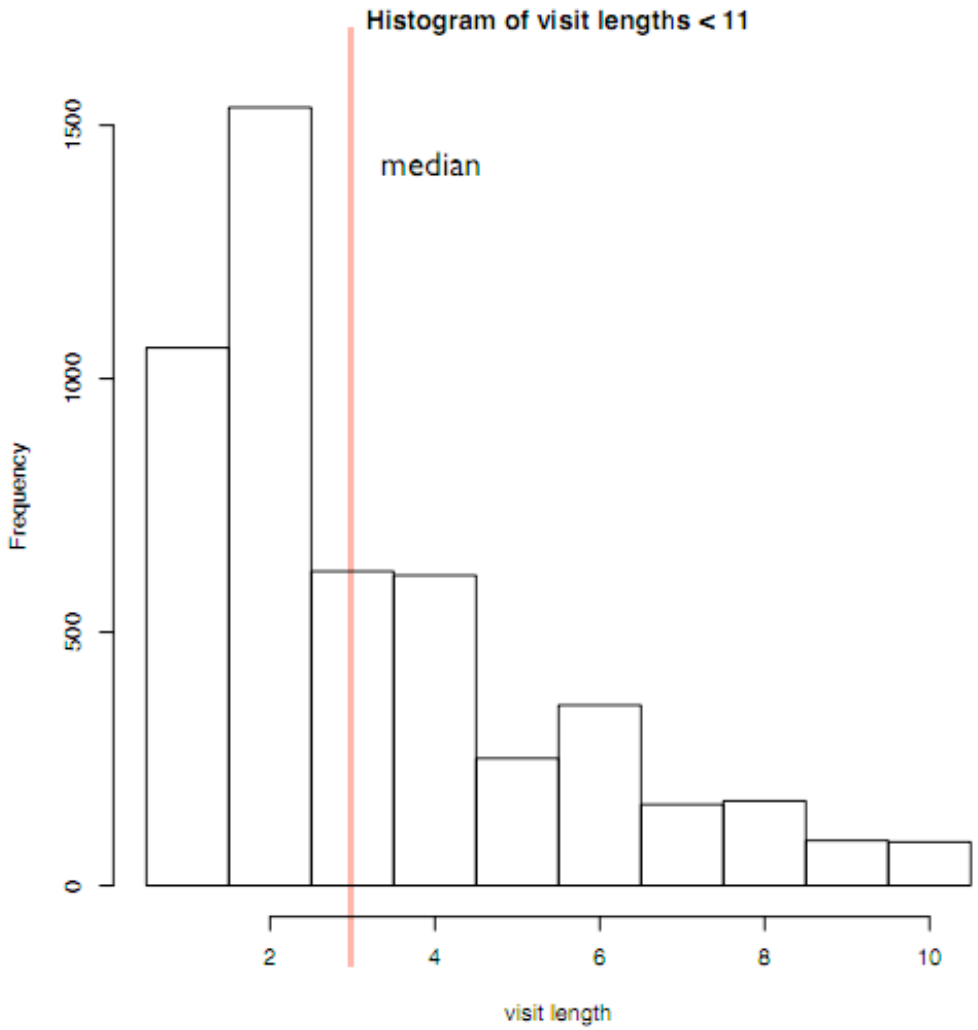


Web visits example

Only 824 of the 5,761 visits are longer than 10 hits

Here we present an even more severe restriction on the x-axis; is this a better way to summarize things?

The red lines again mark the median and mean; um, what's missing?



Example

Interestingly, the mean (11.12) doesn't even appear on a plot that contains (5761-824)/5761 = 86% of the data!

What notion of "center" is this providing, then?

Since a lot of statistical work involves using means, it is often suggested that we **transform the data** to give us something that is "better behaved" or has fewer "extreme" points