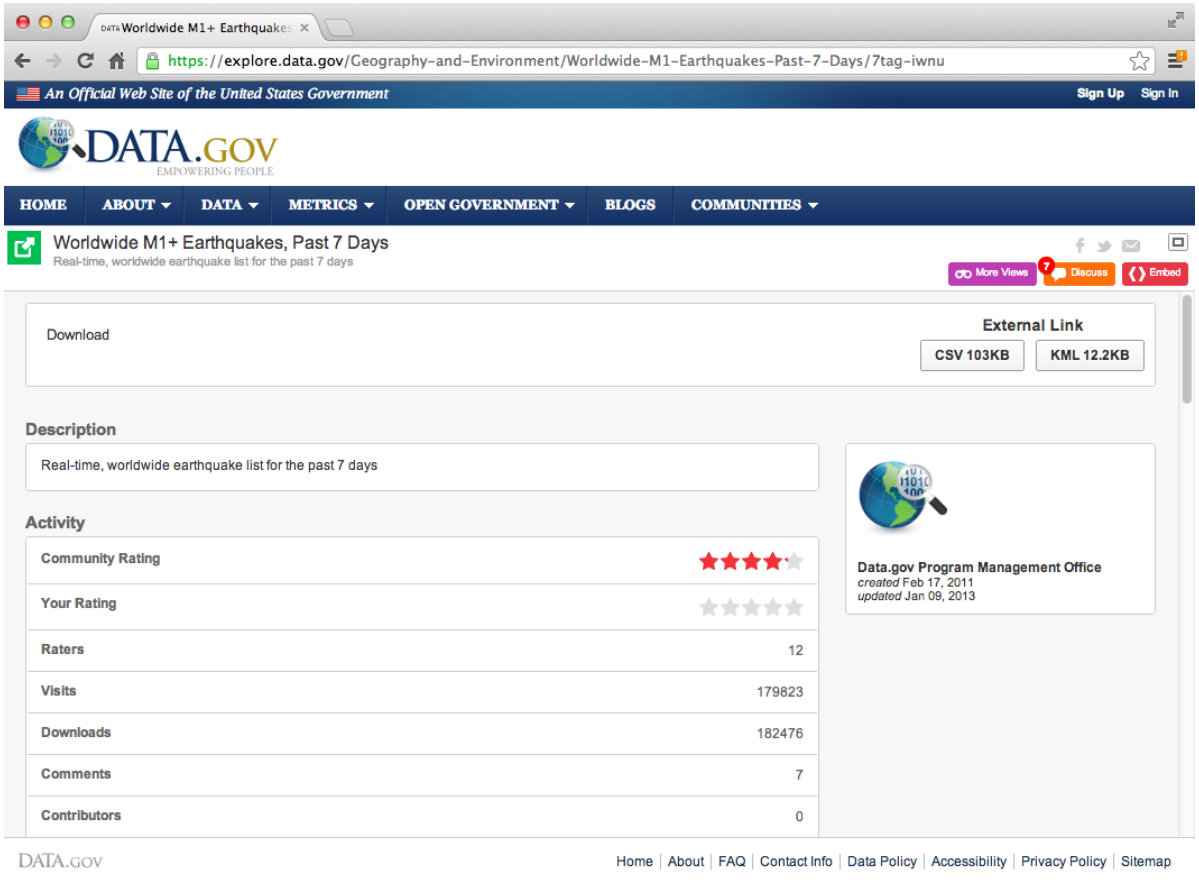
Summarizing data

Jeffrey Leek, Assistant Professor of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Why summarize?

- Data are often too big to look at the whole thing
- The first step in an analysis is to find problems
- When you do these summaries you should be looking for
  - Missing values
  - Values outside of expected ranges
  - Values that seem to be in the wrong units
  - Mislabled variables/columns
  - Variables that are the wrong class

Earthquake data



http://earthquake.usgs.gov/earthquakes/feed/v1.0/

Earthquake data

```
fileUrl <- "http://earthquake.usgs.gov/earthquakes/feed/v1.0/summary/1.0_week.csv"
download.file(fileUrl,destfile="./data/earthquakeData.csv",method="curl")
dateDownloaded <- date()
dateDownloaded
```

```
[1] "Wed Aug 28 14:31:30 2013"
```

```
eData <- read.csv("./data/earthquakeData.csv")
```

Looking at data - the whole thing

```
eData
```

```
                      time latitude longitude   depth  mag magType
1    2013-08-28T18:17:42.700Z   37.405  -121.755   5.500 2.20      Md
2    2013-08-28T17:54:40.400Z   37.404  -121.756   5.600 2.30      Md
3    2013-08-28T17:54:14.000Z   38.781  -122.927   1.700 1.50      Md
4    2013-08-28T17:40:44.860Z   37.025  -117.740   7.000 1.16      ml
5    2013-08-28T16:36:34.778Z   39.669  -119.678  13.813 1.70      ml
6    2013-08-28T16:35:17.000Z   61.603  -141.177   0.100 2.20      Ml
7    2013-08-28T16:12:13.300Z   33.851  -117.811   0.100 1.40      Ml
8    2013-08-28T16:09:26.000Z   62.611  -151.316  86.800 1.20      Ml
9    2013-08-28T16:03:17.100Z   36.438  -121.004   0.000 1.80      Md
10   2013-08-28T15:21:21.200Z   33.517  -116.446  11.500 1.20      Ml
11   2013-08-28T15:13:24.900Z   38.774  -122.715   2.200 1.90      Md
12   2013-08-28T14:39:54.000Z   61.599  -141.230   0.100 2.50      Ml
13   2013-08-28T14:32:21.050Z   39.213    74.585  55.910 4.50      mb
14   2013-08-28T14:26:48.500Z   38.776  -122.716   2.400 2.00      Md
15   2013-08-28T13:58:41.300Z   38.750  -122.701   2.500 1.00      Md
16   2013-08-28T13:25:49.100Z   35.337  -117.914   8.900 1.50      Ml
17   2013-08-28T12:57:40.810Z   42.028    85.850  25.860 4.60      mb
18   2013-08-28T12:50:13.860Z   37.441   144.495  10.110 4.60      mb
19   2013-08-28T12:47:42.000Z   19.367  -155.027   5.500 2.20      Ml
20   2013-08-28T12:44:51.000Z   59.883  -152.292  30.900 1.20      Ml
```

Looking at data - dim(),names(),nrow(),ncol()

```
dim(eData)
```

```
[1] 1076   14
```

```
names(eData)
```

```
 [1] "time"      "latitude"  "longitude" "depth"     "mag"
 [6] "magType"   "nst"       "gap"       "dmin"      "rms"
[11] "net"       "id"        "updated"   "place"
```

```
nrow(eData)
```

```
[1] 1076
```

Looking at the data - quantile()

```
quantile(eData$latitude)
```

```
    0%    25%    50%    75%   100% 
-57.96  33.84  38.82  60.01  67.57 
```

Looking at the data - summary()

```
summary(eData)
```

```
                     time         latitude       longitude
2013-08-21T18:38:20.100Z:   1   Min.   :-58.0   Min.   :-180
2013-08-21T18:38:57.000Z:   1   1st Qu.: 33.8   1st Qu.:-148
2013-08-21T18:55:41.700Z:   1   Median : 38.8   Median :-122
2013-08-21T19:06:43.800Z:   1   Mean   : 41.4   Mean   :-114
2013-08-21T19:11:22.100Z:   1   3rd Qu.: 60.0   3rd Qu.:-116
2013-08-21T19:11:35.900Z:   1   Max.   : 67.6   Max.   : 180
(Other)                  :1070
    depth             mag          magType         nst
Min.   : -2.1   Min.   :1.00   Ml     :633   Min.   :  0.0
1st Qu.:  3.6   1st Qu.:1.20   Md     :310   1st Qu.: 12.0
Median :  9.0   Median :1.60   mb     : 80   Median : 19.0
Mean   : 24.7   Mean   :1.91   ml     : 38   Mean   : 27.5
3rd Qu.: 21.4   3rd Qu.:2.20   ML     :  4   3rd Qu.: 32.0
Max.   :592.3   Max.   :6.10   Mw     :  3   Max.   :249.0
                               (Other):  8   NA's   :519
     gap              dmin            rms             net
Min.   :  0.0   Min.   : 0.0   Min.   :0.010   ak     :369
1st Qu.: 64.8   1st Qu.: 0.0   1st Qu.:0.120   ci     :230
Median : 94.0   Median : 0.1   Median :0.230   nc     :186
Mean   :117.9   Mean   : 0.7   Mean   :0.344   us     : 91
```

Looking at data - class()

```
class(eData)
```

```
[1] "data.frame"
```

```
sapply(eData[1,],class)
```

```
     time  latitude longitude     depth       mag   magType       nst 
 "factor" "numeric" "numeric" "numeric" "numeric"  "factor" "integer" 
      gap      dmin       rms       net        id   updated     place 
"numeric" "numeric" "numeric"  "factor"  "factor"  "factor"  "factor" 
```

Looking at data - unique(), length(), table()

```
unique(eData$net)
```

```
 [1] nc nn ak ci us hv uw pr se uu nm ld at mb
Levels: ak at ci hv ld mb nc nm nn pr se us uu uw
```

```
length(unique(eData$net))
```

```
[1] 14
```

```
table(eData$net)
```

```
 ak  at  ci  hv  ld  mb  nc  nm  nn  pr  se  us  uu  uw
369   1 230  27   2   4 186   6  38  71   4  91  12  35
```

Looking at data - table()

```
table(eData$net,eData$mag)
```

```
    1 1.01 1.03 1.05 1.07 1.1 1.16 1.18 1.2 1.22 1.29 1.3 1.31 1.32
ak 22    0    0    0    0  36    0    0  37    0    0  35    0    0
at  0    0    0    0    0   0    0    0   0    0    0   0    0    0
ci 25    0    0    0    0  28    0    0  30    0    0  25    0    0
hv  2    0    0    0    0   0    0    0   0    0    0   1    0    0
ld  0    0    0    0    0   0    0    0   0    0    0   0    0    0
mb  2    0    0    0    0   0    0    0   0    0    0   1    0    0
nc 24    0    0    0    0  12    0    0  28    0    0  14    0    0
nm  0    0    0    0    0   0    0    0   0    0    0   1    0    0
nn  0    2    1    2    2   0    1    1   0    2    1   0    1    2
pr  0    0    0    0    0   0    0    0   1    0    0   1    0    0
se  0    0    0    0    0   1    0    0   0    0    0   0    0    0
us  0    0    0    0    0   0    0    0   0    0    0   0    0    0
uu  1    0    0    0    0   2    0    0   0    0    0   1    0    0
uw  6    0    0    0    0   3    0    0   3    0    0   1    0    0

   1.39 1.4 1.45 1.47 1.5 1.52 1.58 1.6 1.63 1.65 1.68 1.7 1.78 1.8
ak    0  32    0    0  31    0    0  22    0    0    0  24    0  22
at    0   0    0    0   0    0    0   0    0    0    0   0    0   0
ci    0  37    0    0  27    0    0   9    0    0    0  13    0   7
```

Looking at data - any(), all()

```
eData$latitude[1:10]
```

```
 [1] 37.40 37.40 38.78 37.02 39.67 61.60 33.85 62.61 36.44 33.52
```

```
eData$latitude[1:10] > 40
```

```
 [1] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE
```

```
any(eData$latitude[1:10] > 40)
```

```
[1] TRUE
```

Looking at data - all()

```
eData$latitude[1:10] > 40
```

```
 [1] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE
```

```
all(eData$latitude[1:10] > 40)
```

```
[1] FALSE
```

Looking at subsets - &

```
eData[eData$latitude > 0 & eData$longitude > 0,c("latitude","longitude")]
```

```
    latitude longitude
13    39.213     74.58
17    42.028     85.85
18    37.441    144.50
45    41.827    139.91
107   28.398     99.09
123   32.739     56.41
135   28.252     99.33
163   37.595    142.05
181   44.506    149.06
205   47.122    152.66
208    2.171    128.61
247   56.546    112.70
269   34.024     87.94
302   33.176     94.10
316   30.038     97.88
345   34.382    141.01
355   49.577    155.53
375   38.826     69.93
431   22.145    144.04
449   35.428    140.16
```

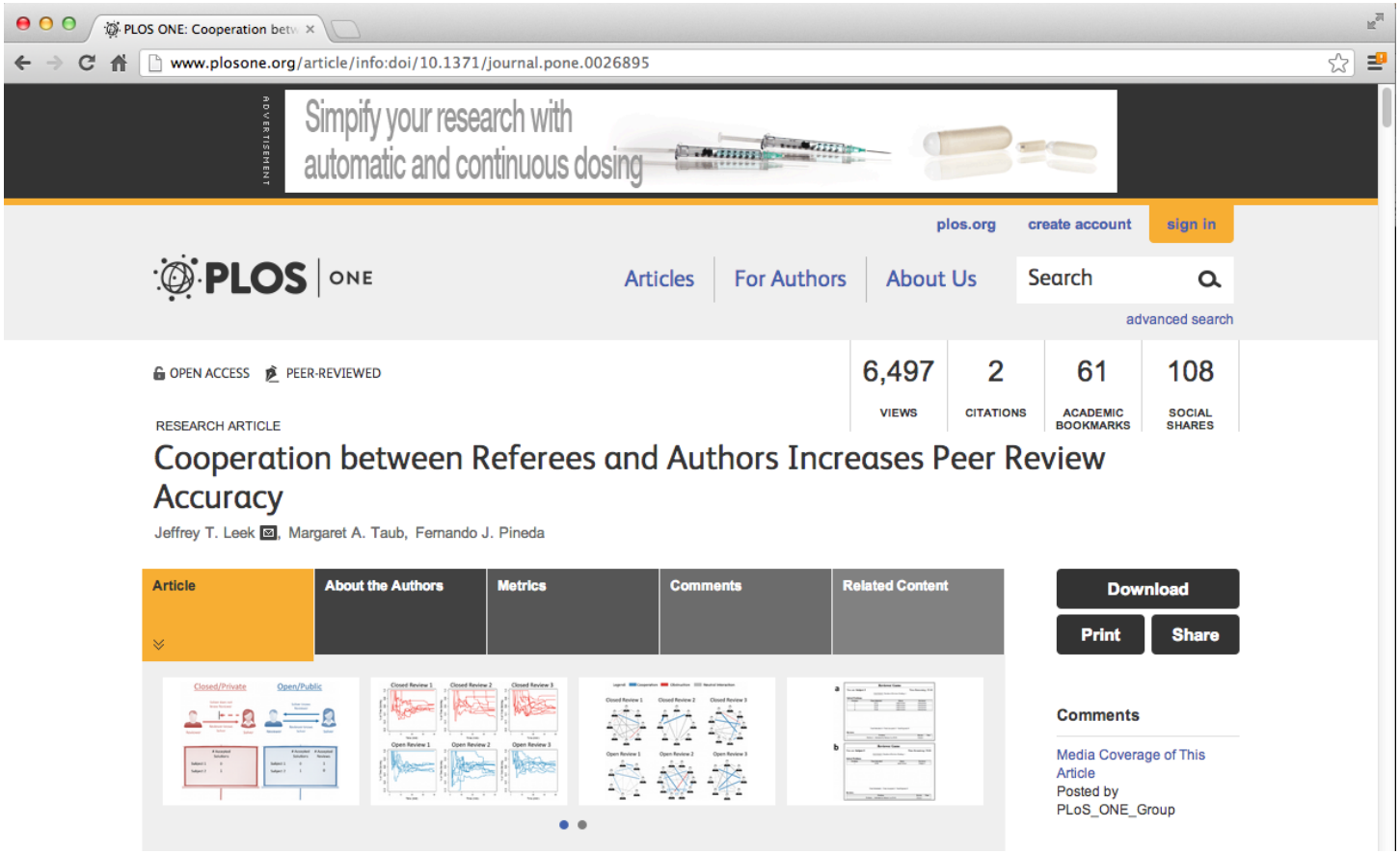
Looking at subsets - |

```
eData[eData$latitude > 0 | eData$longitude > 0,c("latitude","longitude")]
```

```
     latitude longitude
1      37.405  -121.755
2      37.404  -121.756
3      38.781  -122.927
4      37.025  -117.740
5      39.669  -119.678
6      61.603  -141.177
7      33.851  -117.811
8      62.611  -151.316
9      36.438  -121.004
10     33.517  -116.446
11     38.774  -122.715
12     61.599  -141.230
13     39.213    74.585
14     38.776  -122.716
15     38.750  -122.701
16     35.337  -117.914
17     42.028    85.850
18     37.441   144.495
19     19.367  -155.027
20     59.883  -152.292
```

Peer review experiment data

- Data on submissions/reviews in an experiment



http://www.plosone.org/article/info:doi/10.1371/journal.pone.0026895

Peer review data

```
fileUrl1 <- "https://dl.dropbox.com/u/7710864/data/reviews-apr29.csv"
fileUrl2 <- "https://dl.dropbox.com/u/7710864/data/solutions-apr29.csv"
download.file(fileUrl1,destfile="./data/reviews.csv",method="curl")
download.file(fileUrl2,destfile="./data/solutions.csv",method="curl")
reviews <- read.csv("./data/reviews.csv"); solutions <- read.csv("./data/solutions.csv")
head(reviews,2)
```

```
                              X.html.
1   <head><title>Found</title></head>
2                              <body>
```

```
head(solutions,2)
```

```
                              X.html.
1   <head><title>Found</title></head>
2                              <body>
```

Find if there are missing values - is.na()

```
is.na(reviews$time_left[1:10])
```

```
logical(0)
```

```
sum(is.na(reviews$time_left))
```

```
[1] 0
```

```
table(is.na(reviews$time_left))
```

```
< table of extent 0 >
```

Important table()/NA issue

```
table(c(0,1,2,3,NA,3,3,2,2,3))
```

```

0 1 2 3
1 1 3 4
```

```
table(c(0,1,2,3,NA,3,3,2,2,3),useNA="ifany")
```

```

   0    1    2    3 <NA>
   1    1    3    4    1
```

Summarizing columns/rows - rowSums(),rowMeans(),colSums(),colMeans()

- Important parameters: *x*, *na.rm*

```
colSums(reviews)
```

```
Error: 'x' must be numeric
```

Summarizing columns/rows - rowSums(),rowMeans(),colSums(),colMeans()

```
colMeans(reviews,na.rm=TRUE)
```

```
Error: 'x' must be numeric
```

```
rowMeans(reviews,na.rm=TRUE)
```

```
Error: 'x' must be numeric
```