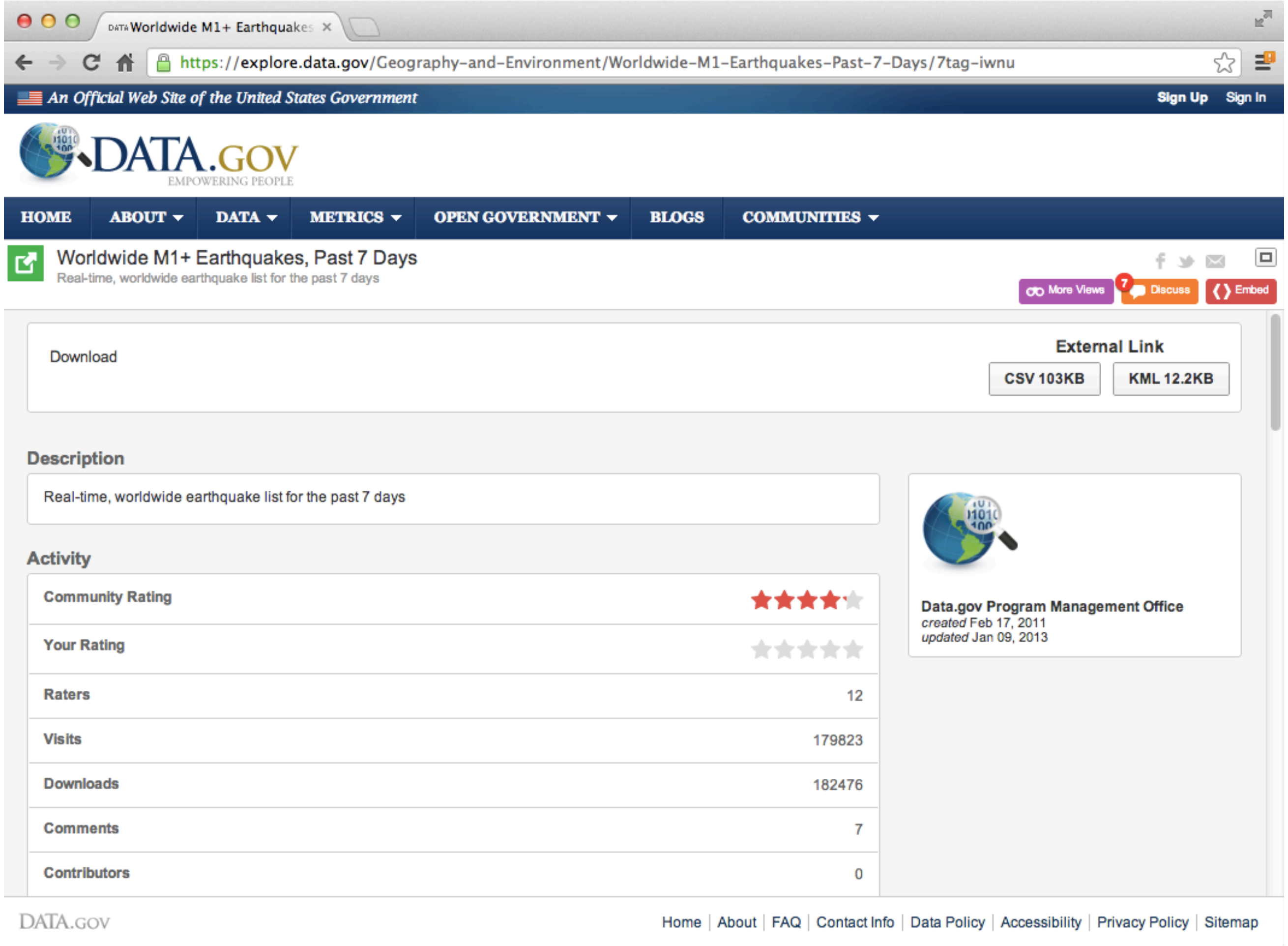
Summarizing data

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


Worldwide M1+ Earthquakes
https://explore.data.gov/Geography-and-Environment/Worldwide-M1-Earthquakes-Past-7-Days/7tag-iwnu
An Official Web Site of the United States Government
Sign Up
Sign In
DATA.GOV
EMPOWERING PEOPLE
HOME
ABOUT
DATA
METRICS
OPEN GOVERNMENT
BLOGS
COMMUNITIES
More Views
7
Discuss
Embed
DATA.GOV
Home | About | FAQ | Contact Info | Data Policy | Accessibility | Privacy Policy | Sitemap

Worldwide M1+ Earthquakes, Past 7 Days

Real-time, worldwide earthquake list for the past 7 days

Download

External Link

CSV 103KB | KML 12.2KB

Description

Real-time, worldwide earthquake list for the past 7 days

Activity

Community Rating	★★★★☆
Your Rating	☆☆☆☆☆
Raters	12
Visits	179823
Downloads	182476
Comments	7
Contributors	0

Data.gov Program Management Office
created Feb 17, 2011
updated Jan 09, 2013

https://explore.data.gov/Geography-and-Environment/Worldwide-M1-Earthquakes-Past-7-Days/7tag-iwnu

Earthquake data

```
fileUrl <- "http://earthquake.usgs.gov/earthquakes/catalogs/eqs7day-M1.txt"
download.file(fileUrl,destfile="./data/earthquakeData.csv",method="curl")
dateDownloaded <- date()
dateDownloaded
```

```
[1] "Sun Jan 27 00:23:22 2013"
```

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

Looking at data - the whole thing

```
eData
```

```
     Src     Eqid Version                            Datetime
1     nc 71929481       1   Sunday, January 27, 2013 05:03:01 UTC
2     ci 15278017       0   Sunday, January 27, 2013 04:59:04 UTC
3     ak 10645573       1   Sunday, January 27, 2013 04:55:09 UTC
4     nc 71929476       0   Sunday, January 27, 2013 04:51:48 UTC
5     nn 00401016       9   Sunday, January 27, 2013 04:45:19 UTC
6     ak 10645564       1   Sunday, January 27, 2013 04:16:45 UTC
7     hv 60459531       2   Sunday, January 27, 2013 04:15:57 UTC
8     ak 10645555       1   Sunday, January 27, 2013 04:14:35 UTC
9     ci 15278009       0   Sunday, January 27, 2013 04:07:44 UTC
10    us c000ewb3       7   Sunday, January 27, 2013 04:05:42 UTC
11    ci 15278001       0   Sunday, January 27, 2013 03:54:27 UTC
12    hv 60459521       1   Sunday, January 27, 2013 03:50:13 UTC
13    hv 60459516       2   Sunday, January 27, 2013 03:43:56 UTC
14    ak 10645533       1   Sunday, January 27, 2013 03:25:17 UTC
15    ak 10645528       1   Sunday, January 27, 2013 03:18:17 UTC
16    us c000ewax       6   Sunday, January 27, 2013 03:17:57 UTC
17    ci 15277993       0   Sunday, January 27, 2013 02:47:04 UTC
```

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

```
dim(eData)
```

```
[1] 1057   10
```

```
names(eData)
```

```
 [1] "Src"       "Eqid"      "Version"   "Datetime"  "Lat"      
 [6] "Lon"       "Magnitude" "Depth"     "NST"       "Region"   
```

```
nrow(eData)
```

```
[1] 1057
```

Looking at the data - quantile(),summary()

```
quantile(eData$Lat)
```

```
    0%    25%    50%    75%   100% 
-61.30  35.56  38.77  52.58  67.66 
```

```
summary(eData)
```

```
      Src            Eqid         Version   
 ak     :330   00400150:   1   2      :379  
 nc     :247   00400153:   1   0      :195  
 ci     :145   00400155:   1   1      :168  
 nn     : 92   00400156:   1   9      : 97  
 us     : 89   00400157:   1   3      : 82  
 pr     : 40   00400159:   1   4      : 43  
 (Other):114   (Other) :1051   (Other): 93  
                                  Datetime         Lat       
 Monday, January 21, 2013 11:00:00 UTC:   2   Min.   :-61.3  
 Friday, January 25, 2013 00:06:25 UTC:   1   1st Qu.: 35.6  
```

Looking at data - class()

```
class(eData)
```

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

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

```
      Src      Eqid   Version  Datetime       Lat       Lon Magnitude 
 "factor"  "factor"  "factor"  "factor" "numeric" "numeric" "numeric" 
    Depth       NST    Region 
"numeric" "integer"  "factor" 
```

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

```
unique(eData$Src)
```

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

```
length(unique(eData$Src))
```

```
[1] 11
```

```
table(eData$Src)
```

```
 ak  ci  hv  mb  nc  nm  nn  pr  us  uu  uw 
330 145  29  10 247   2  92  40  89  40  33 
```

Looking at data - table()

```
table(eData$Src,eData$Version)
```

```
      0   1   2   3   4   5   6   7   8   9   A   B   D   E
 ak   0  93 211  26   0   0   0   0   0   0   0   0   0   0
 ci  64   0  67   7   3   3   1   0   0   0   0   0   0   0
 hv   0  14  11   0   2   2   0   0   0   0   0   0   0   0
 mb   0   0  10   0   0   0   0   0   0   0   0   0   0   0
 nc  91  46  51  37  10   4   3   1   1   1   1   1   0   0
 nm   0   0   0   0   0   0   0   0   0   0   2   0   0   0
 nn   0   0   0   0   0   0   0   0   0  92   0   0   0   0
 pr  40   0   0   0   0   0   0   0   0   0   0   0   0   0
 us   0   0   2   0  14  13  24  13  11   4   4   2   1   1
 uu   0   0  15   6  14   3   2   0   0   0   0   0   0   0
 uw   0  15  12   6   0   0   0   0   0   0   0   0   0   0
```

Looking at data - any(), all()

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

```
 [1] 38.83 36.04 65.23 39.56 37.26 62.10 19.41 63.51 32.91 -5.17
```

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

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

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

```
[1] TRUE
```

Looking at data - all()

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

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

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

```
[1] FALSE
```

Looking at subsets - &

```
eData[eData$Lat > 0 & eData$Lon > 0,c("Lat","Lon")]
```

```
       Lat    Lon
51   5.486 127.05
56  39.749  77.30
58  38.295  46.81
110 34.571  24.10
129 51.130 179.35
134  9.438 126.10
146 38.426  73.36
153 49.728 155.69
155 43.337  18.77
160 29.379 132.20
175 44.280  10.53
193 31.763  50.95
239  4.998  95.96
325 53.564 142.75
348 38.608  73.49
359 27.771  56.41
385 49.825  87.60
```

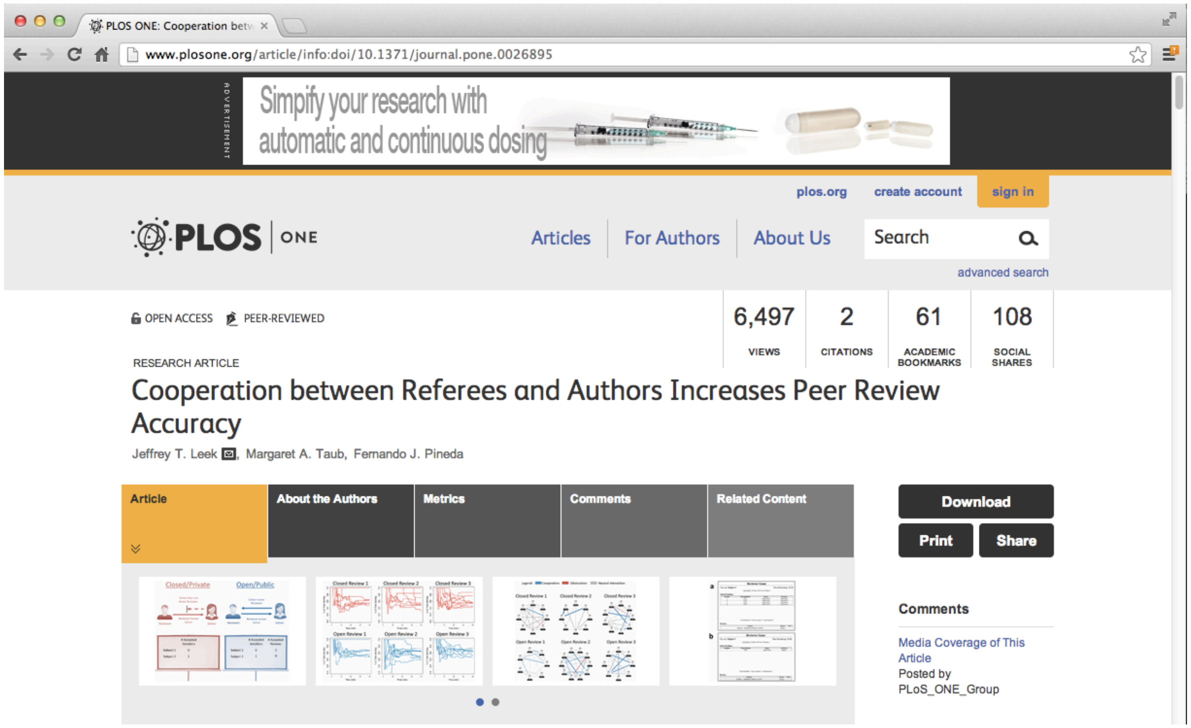
Looking at subsets - |

```
eData[eData$Lat > 0 | eData$Lon > 0,c("Lat","Lon")]
```

```
          Lat     Lon
1     38.8292 -122.81
2     36.0403 -117.35
3     65.2271 -149.51
4     39.5573 -121.99
5     37.2587 -114.07
6     62.1046 -150.70
7     19.4065 -155.26
8     63.5132 -150.83
9     32.9112 -116.25
10    -5.1704  102.94
11    35.5633 -118.53
12    19.2960 -155.38
13    19.9262 -155.54
14    62.1638 -149.58
15    63.2917 -149.24
16    34.2925 -106.71
17    33.6293 -116.69
```

Peer review experiment data

- Data on submissions/reviews in an experiment



6,497	2	61	108
VIEWS	CITATIONS	ACADEMIC BOOKMARKS	SOCIAL SHARES

RESEARCH ARTICLE

Cooperation between Referees and Authors Increases Peer Review Accuracy

Jeffrey T. Leek, Margaret A. Taub, Fernando J. Pineda

Comments

Media Coverage of This Article
Posted by PLoS_ONE_Group

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)
```

```
  id solution_id reviewer_id      start       stop time_left accept
1  1           3          27 1304095698 1304095758      1754      1
2  2           4          22 1304095188 1304095206      2306      1
```

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

```
  id problem_id subject_id      start       stop time_left answer
1  1        156         29 1304095119 1304095169      2343      B
2  2        269         25 1304095119 1304095183      2329      C
```

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

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

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

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

```
[1] 84
```

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

```

FALSE  TRUE 
  115    84 
```

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)
```

```
         id solution_id reviewer_id       start        stop
      19900       19929        5064          NA          NA
  time_left      accept
         NA          NA
```

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

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

```
         id solution_id reviewer_id       start        stop 
  1.000e+02   1.001e+02   2.545e+01   1.304e+09   1.304e+09 
  time_left      accept 
  1.114e+03   6.435e-01 
```

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

```
  [1] 3.726e+08 3.726e+08 3.726e+08 3.726e+08 3.726e+08 3.726e+08
  [7] 3.726e+08 1.300e+01 3.726e+08 3.726e+08 3.726e+08 3.726e+08
 [13] 3.726e+08 3.726e+08 3.726e+08 3.726e+08 1.967e+01 3.726e+08
 [19] 3.726e+08 1.933e+01 3.726e+08 3.726e+08 3.726e+08 2.433e+01
 [25] 2.367e+01 2.367e+01 3.726e+08 3.726e+08 3.726e+08 3.726e+08
 [31] 3.726e+08 3.726e+08 3.726e+08 3.726e+08 3.133e+01 3.726e+08
 [37] 3.267e+01 3.726e+08 3.400e+01 3.726e+08 3.200e+01 3.726e+08
```