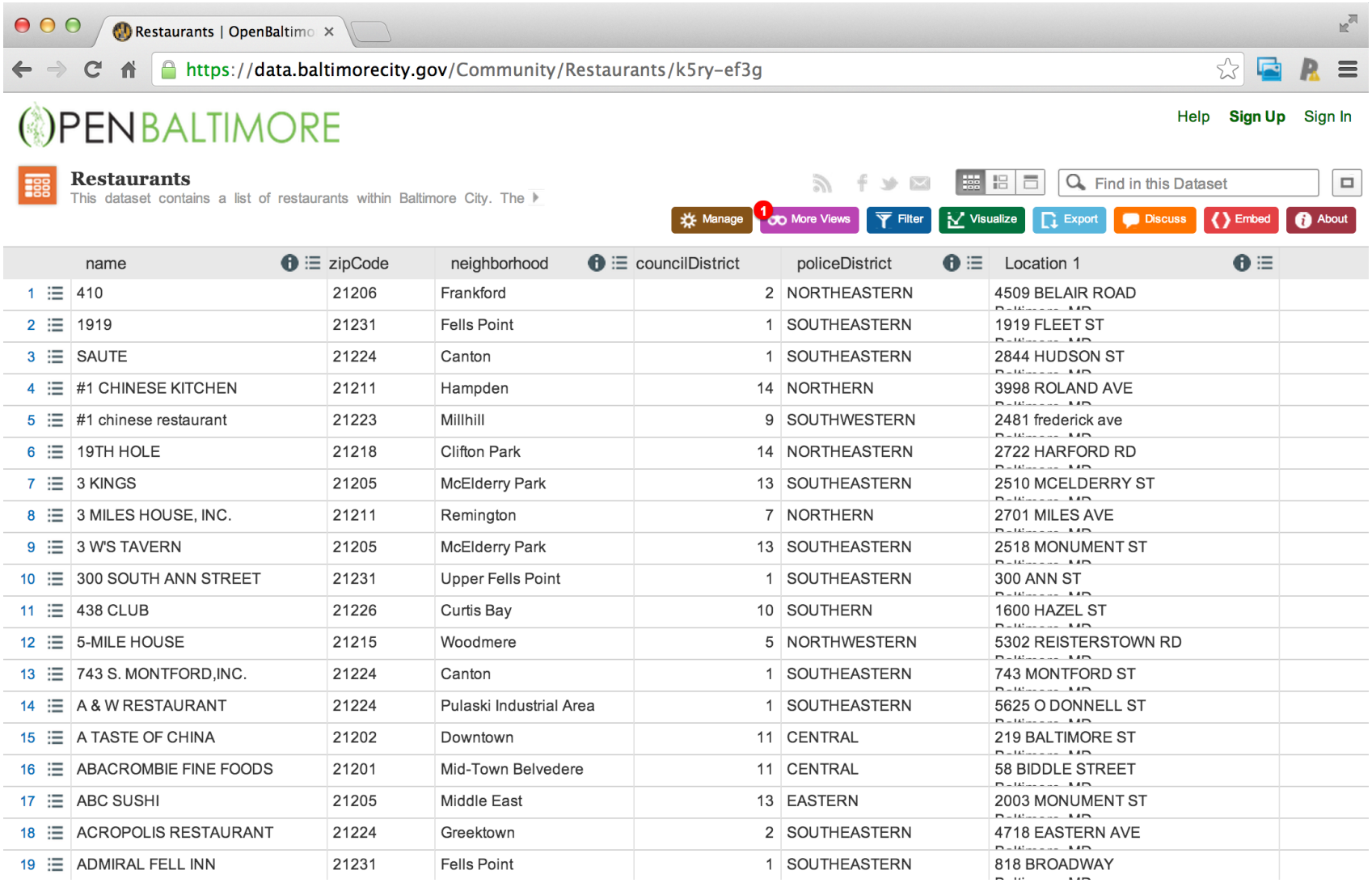
Summarizing data

Jeffrey Leek

May 18, 2016

Example data set



	name	zipCode	neighborhood	councilDistrict	policeDistrict	Location 1
1	410	21206	Frankford	2	NORTHEASTERN	4509 BELAIR ROAD
2	1919	21231	Fells Point	1	SOUTHEASTERN	1919 FLEET ST
3	SAUTE	21224	Canton	1	SOUTHEASTERN	2844 HUDSON ST
4	#1 CHINESE KITCHEN	21211	Hampden	14	NORTHERN	3998 ROLAND AVE
5	#1 chinese restaurant	21223	Millhill	9	SOUTHWESTERN	2481 frederick ave
6	19TH HOLE	21218	Clifton Park	14	NORTHEASTERN	2722 HARFORD RD
7	3 KINGS	21205	McElderry Park	13	SOUTHEASTERN	2510 MCELDERRY ST
8	3 MILES HOUSE, INC.	21211	Remington	7	NORTHERN	2701 MILES AVE
9	3 W'S TAVERN	21205	McElderry Park	13	SOUTHEASTERN	2518 MONUMENT ST
10	300 SOUTH ANN STREET	21231	Upper Fells Point	1	SOUTHEASTERN	300 ANN ST
11	438 CLUB	21226	Curtis Bay	10	SOUTHERN	1600 HAZEL ST
12	5-MILE HOUSE	21215	Woodmere	5	NORTHWESTERN	5302 REISTERSTOWN RD
13	743 S. MONTFORD,INC.	21224	Canton	1	SOUTHEASTERN	743 MONTFORD ST
14	A & W RESTAURANT	21224	Pulaski Industrial Area	1	SOUTHEASTERN	5625 O DONNELL ST
15	A TASTE OF CHINA	21202	Downtown	11	CENTRAL	219 BALTIMORE ST
16	ABACROMBIE FINE FOODS	21201	Mid-Town Belvedere	11	CENTRAL	58 BIDDLE STREET
17	ABC SUSHI	21205	Middle East	13	EASTERN	2003 MONUMENT ST
18	ACROPOLIS RESTAURANT	21224	Greektown	2	SOUTHEASTERN	4718 EASTERN AVE
19	ADMIRAL FELL INN	21231	Fells Point	1	SOUTHEASTERN	818 BROADWAY

https://data.baltimorecity.gov/Community/Restaurants/k5ry-ef3g

Getting the data from the web

```
if(!file.exists("./data")){dir.create("./data")}
fileUrl <- "https://data.baltimorecity.gov/api/views/k5ry-e
download.file(fileUrl,destfile="./data/restaurants.csv",met
restData <- read.csv("./data/restaurants.csv")
```

Look at a bit of the data

```r
head(restData,n=3)
```

```
##    name zipCode neighborhood councilDistrict policeDistr
## 1   410   21206    Frankford               2   NORTHEAST
## 2  1919   21231  Fells Point               1   SOUTHEAST
## 3 SAUTE   21224       Canton               1   SOUTHEAST
##                          Location.1
## 1 4509 BELAIR ROAD\nBaltimore, MD\n
## 2    1919 FLEET ST\nBaltimore, MD\n
## 3   2844 HUDSON ST\nBaltimore, MD\n
```

```r
tail(restData,n=3)
```

```
##                  name zipCode  neighborhood councilDistr
## 1325 ZINK'S CAF\u0090   21213 Belair-Edison
## 1326     ZISSIMOS BAR   21211       Hampden
## 1327           ZORBAS   21224     Greektown
##                          Location.1
```

Make summary

```r
summary(restData)
```

```
##                                  name          zipCode
##  MCDONALD'S                        :   8   Min.   :-21226   Do
##  POPEYES FAMOUS FRIED CHICKEN:   7   1st Qu.: 21202   Fe
##  SUBWAY                            :   6   Median : 21218   In
##  KENTUCKY FRIED CHICKEN            :   5   Mean   : 21185   Ca
##  BURGER KING                       :   4   3rd Qu.: 21226   Fe
##  DUNKIN DONUTS                     :   4   Max.   : 21287   Mo
##  (Other)                           :1293                     ((
##  councilDistrict       policeDistrict
##  Min.   : 1.000   SOUTHEASTERN:385
##  1st Qu.: 2.000   CENTRAL      :288
##  Median : 9.000   SOUTHERN     :213
##  Mean   : 7.191   NORTHERN     :157
##  3rd Qu.:11.000   NORTHEASTERN: 72
##  Max.   :14.000   EASTERN      : 67
##                   (Other)      :145
```

More in depth information

```r
str(restData)
```

```
## 'data.frame':    1327 obs. of  6 variables:
##  $ name           : Factor w/ 1277 levels "#1 CHINESE K
##  $ zipCode        : int  21206 21231 21224 21211 21223 2
##  $ neighborhood   : Factor w/ 173 levels "Abell","Arling
##  $ councilDistrict: int  2 1 1 14 9 14 13 7 13 1 ...
##  $ policeDistrict : Factor w/ 9 levels "CENTRAL","EASTER
##  $ Location.1     : Factor w/ 1210 levels "1 BIDDLE ST\n
```

Quantiles of quantitative variables

```r
quantile(restData$councilDistrict,na.rm=TRUE)
```

```
##   0%  25%  50%  75% 100%
##    1    2    9   11   14
```

```r
quantile(restData$councilDistrict,probs=c(0.5,0.75,0.9))
```

```
## 50% 75% 90%
##   9  11  12
```

Make table

```r
table(restData$zipCode,useNA="ifany")
```

```
##
## -21226  21201  21202  21205  21206  21207  21208  21209
##      1    136    201     27     30      4      1      8
##  21212  21213  21214  21215  21216  21217  21218  21220
##     28     31     17     54     10     32     69      1
##  21224  21225  21226  21227  21229  21230  21231  21234
##    199     19     18      4     13    156    127      7
##  21251  21287
##      2      1
```

Make table

```r
table(restData$councilDistrict,restData$zipCode)
```

```
##
##       -21226 21201 21202 21205 21206 21207 21208 21209 21
##   1        0     0    37     0     0     0     0     0
##   2        0     0     0     3    27     0     0     0
##   3        0     0     0     0     0     0     0     0
##   4        0     0     0     0     0     0     0     0
##   5        0     0     0     0     0     3     0     6
##   6        0     0     0     0     0     0     0     1
##   7        0     0     0     0     0     0     0     1
##   8        0     0     0     0     0     1     0     0
##   9        0     1     0     0     0     0     0     0
##   10       1     0     1     0     0     0     0     0
##   11       0   115   139     0     0     0     1     0
##   12       0    20    24     4     0     0     0     0
##   13       0     0     0    20     3     0     0     0
##   14       0     0     0     0     0     0     0     0
```

Check for missing values

```r
sum(is.na(restData$councilDistrict))
```

```
## [1] 0
```

```r
any(is.na(restData$councilDistrict))
```

```
## [1] FALSE
```

```r
all(restData$zipCode > 0)
```

```
## [1] FALSE
```

Row and column sums

```r
colSums(is.na(restData))
```

```
##              name           zipCode     neighborhood councilD
##                 0                 0                0
##   policeDistrict       Location.1
##                 0                 0
```

```r
all(colSums(is.na(restData))==0)
```

```
## [1] TRUE
```

Values with specific characteristics

```r
table(restData$zipCode %in% c("21212"))
```

```
##
## FALSE  TRUE
##  1299    28
```

```r
table(restData$zipCode %in% c("21212","21213"))
```

```
##
## FALSE  TRUE
##  1268    59
```

Values with specific characteristics

```
restData[restData$zipCode %in% c("21212","21213"),]
```

```
##                                        name zipCode
## 29                        BAY ATLANTIC CLUB   21212
## 39                              BERMUDA BAR   21213
## 92                                ATWATER'S   21212
## 111             BALTIMORE ESTONIAN SOCIETY   21213
## 187                                CAFE ZEN   21212
## 220                   CERIELLO FINE FOODS   21212
## 266    CLIFTON PARK GOLF COURSE SNACK BAR   21213
## 276                CLUB HOUSE BAR & GRILL   21213
## 289                 CLUBHOUSE BAR & GRILL   21213
## 291                              COCKY LOU'S   21213
## 362       DREAM TAVERN, CARRIBEAN  U.S.A.   21213
## 373                           DUNKIN DONUTS   21212
## 383        EASTSIDE  SPORTS  SOCIAL  CLUB   21213
## 417                       FIELDS OLD TRAIL   21212
## 475                                GRAND CRU   21212
```

Cross tabs

```r
data(UCBAdmissions)
DF = as.data.frame(UCBAdmissions)
summary(DF)
```

```
##       Admit       Gender   Dept       Freq
##  Admitted:12   Male  :12   A:4   Min.   :  8.0
##  Rejected:12   Female:12   B:4   1st Qu.: 80.0
##                            C:4   Median :170.0
##                            D:4   Mean   :188.6
##                            E:4   3rd Qu.:302.5
##                            F:4   Max.   :512.0
```

Cross tabs

```r
xt <- xtabs(Freq ~ Gender + Admit,data=DF)
xt
```

```
##          Admit
## Gender    Admitted Rejected
##   Male        1198     1493
##   Female       557     1278
```

Flat tables

```r
warpbreaks$replicate <- rep(1:9, len = 54)
xt = xtabs(breaks ~.,data=warpbreaks)
xt
```

```
## , , replicate = 1
##
##     tension
## wool  L  M  H
##    A 26 18 36
##    B 27 42 20
##
## , , replicate = 2
##
##     tension
## wool  L  M  H
##    A 30 21 21
##    B 14 26 21
##
```

Flat tables

```r
ftable(xt)
```

```
##                 replicate  1  2  3  4  5  6  7  8  9
## wool tension
## A    L                    26 30 54 25 70 52 51 26 67
##      M                    18 21 29 17 12 18 35 30 36
##      H                    36 21 24 18 10 43 28 15 26
## B    L                    27 14 29 19 29 31 41 20 44
##      M                    42 26 19 16 39 28 21 39 29
##      H                    20 21 24 17 13 15 15 16 28
```

Size of a data set

```r
fakeData = rnorm(1e5)
object.size(fakeData)
```

```
## 800040 bytes
```

```r
print(object.size(fakeData),units="Mb")
```

```
## 0.8 Mb
```