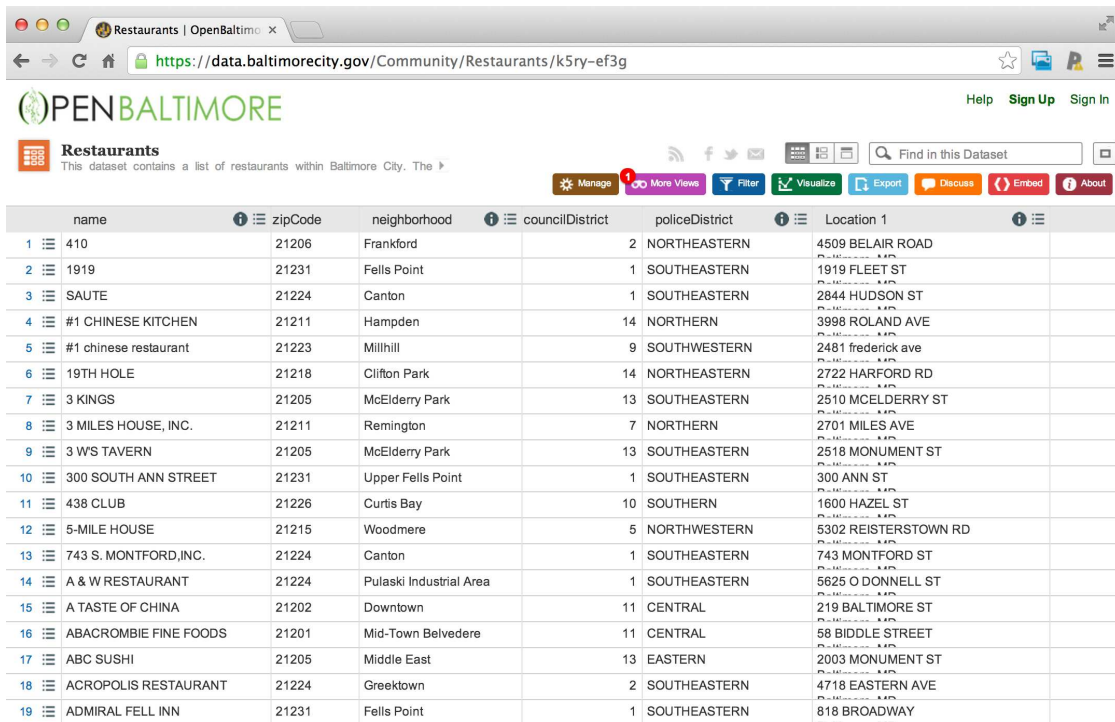




Summarizing data

Jeffrey Leek
Johns Hopkins Bloomberg School of Public Health

Example data set



Restaurants
This dataset contains a list of restaurants within Baltimore City. The ▶

Manage More Views Filter Visualize Export Discuss Embed About

	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 WVS 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-ef3g/rows.csv?accessType=DOWNLOAD"  
download.file(fileUrl,destfile="./data/restaurants.csv",method="curl")  
restData <- read.csv("./data/restaurants.csv")
```

Look at a bit of the data

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

	name	zipCode	neighborhood	councilDistrict	policeDistrict	Location.1
1	410	21206	Frankford	2	NORTHEASTERN	4509 BELAIR ROAD\nBaltimore, MD\n
2	1919	21231	Fells Point	1	SOUTHEASTERN	1919 FLEET ST\nBaltimore, MD\n
3	SAUTE	21224	Canton	1	SOUTHEASTERN	2844 HUDSON ST\nBaltimore, MD\n

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

	name	zipCode	neighborhood	councilDistrict	policeDistrict	Location.1
1325	ZINK'S CAF\u0090	21213	Belair-Edison	13	NORTHEASTERN	
1326	ZISSIMOS BAR	21211	Hampden	7	NORTHERN	
1327	ZORBAS	21224	Greektown	2	SOUTHEASTERN	
						Location.1
1325	3300 LAWNVIEW AVE\n					Baltimore, MD\n
1326	1023 36TH ST\n					Baltimore, MD\n
1327	4710 EASTERN Ave\n					Baltimore, MD\n

Make summary

```
summary(restData)
```

		name	zipCode	neighborhood	councilDistrict
MCDONALD'S	: 8	Min.	: -21226	Downtown	: 128 Min. : 1.00
POPEYES FAMOUS FRIED CHICKEN:	7	1st Qu.:	21202	Fells Point	: 91 1st Qu.: 2.00
SUBWAY	: 6	Median	: 21218	Inner Harbor:	89 Median : 9.00
KENTUCKY FRIED CHICKEN	: 5	Mean	: 21185	Canton	: 81 Mean : 7.19
BURGER KING	: 4	3rd Qu.:	21226	Federal Hill:	42 3rd Qu.: 11.00
DUNKIN DONUTS	: 4	Max.	: 21287	Mount Vernon:	33 Max. : 14.00
(Other)	: 1293			(Other)	: 863
		policeDistrict	Location.1		
SOUTHEASTERN:	385	1101 RUSSELL ST\n	Baltimore, MD\n:	9	
CENTRAL	: 288	201 PRATT ST\n	Baltimore, MD\n:	8	
SOUTHERN	: 213	2400 BOSTON ST\n	Baltimore, MD\n:	8	
NORTHERN	: 157	300 LIGHT ST\n	Baltimore, MD\n:	5	
NORTHEASTERN:	72	300 CHARLES ST\n	Baltimore, MD\n:	4	
EASTERN	: 67	301 LIGHT ST\n	Baltimore, MD\n:	4	
(Other)	: 145	(Other)		: 1289	

Mpre in depth information

```
str(restData)
```

```
'data.frame':  1327 obs. of  6 variables:
 $ name          : Factor w/ 1277 levels "#1 CHINESE KITCHEN",...: 9 3 992 1 2 4 5 6 7 8 ...
 $ zipCode       : int  21206 21231 21224 21211 21223 21218 21205 21211 21205 21231 ...
 $ neighborhood  : Factor w/ 173 levels "Abell","Arlington",...: 53 52 18 66 104 33 98 133 98 157 ...
 $ councilDistrict: int  2 1 1 14 9 14 13 7 13 1 ...
 $ policeDistrict : Factor w/ 9 levels "CENTRAL","EASTERN",...: 3 6 6 4 8 3 6 4 6 6 ...
 $ Location.1    : Factor w/ 1210 levels "1 BIDDLE ST\nBaltimore, MD\n",...: 835 334 554 755 492 537 5
```

Quantiles of quantitative variables

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

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

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

50%	75%	90%
9	11	12

Make table

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

-21226	21201	21202	21205	21206	21207	21208	21209	21210	21211	21212	21213	21214	21215
1	136	201	27	30	4	1	8	23	41	28	31	17	54
21216	21217	21218	21220	21222	21223	21224	21225	21226	21227	21229	21230	21231	21234
10	32	69	1	7	56	199	19	18	4	13	156	127	7
21237	21239	21251	21287										
1	3	2	1										

Make table

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

	-21226	21201	21202	21205	21206	21207	21208	21209	21210	21211	21212	21213	21214	21215	21216
1	0	0	37	0	0	0	0	0	0	0	0	2	0	0	0
2	0	0	0	3	27	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	2	17	0	0
4	0	0	0	0	0	0	0	0	0	0	27	0	0	0	0
5	0	0	0	0	0	3	0	6	0	0	0	0	0	31	0
6	0	0	0	0	0	0	0	1	19	0	0	0	0	15	1
7	0	0	0	0	0	0	0	1	0	27	0	0	0	6	7
8	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
9	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2
10	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
11	0	115	139	0	0	0	1	0	0	0	1	0	0	0	0
12	0	20	24	4	0	0	0	0	0	0	0	13	0	0	0
13	0	0	0	20	3	0	0	0	0	0	0	13	0	1	0
14	0	0	0	0	0	0	0	0	4	14	0	1	0	1	0

21217 21218 21220 21222 21223 21224 21225 21226 21227 21229 21230 21231 21234 21237 21239

Check for missing values

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

```
[1] 0
```

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

```
[1] FALSE
```

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

```
[1] FALSE
```

Row and column sums

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

name	zipCode	neighborhood	councilDistrict	policeDistrict	Location.1
0	0	0	0	0	0

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

```
[1] TRUE
```

Values with specific characteristics

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

FALSE	TRUE
1299	28

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

FALSE	TRUE
1268	59

Values with specific characteristics

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

	name	zipCode	neighborhood	councilDistrict
29	BAY ATLANTIC CLUB	21212	Downtown	11
39	BERMUDA BAR	21213	Broadway East	12
92	ATWATER'S	21212	Chinquapin Park-Belvedere	4
111	BALTIMORE ESTONIAN SOCIETY	21213	South Clifton Park	12
187	CAFE ZEN	21212	Rosebank	4
220	CERIELLO FINE FOODS	21212	Chinquapin Park-Belvedere	4
266	CLIFTON PARK GOLF COURSE SNACK BAR	21213	Darley Park	14
276	CLUB HOUSE BAR & GRILL	21213	Orangeville Industrial Area	13
289	CLUBHOUSE BAR & GRILL	21213	Orangeville Industrial Area	13
291	COCKY LOU'S	21213	Broadway East	12
362	DREAM TAVERN, CARRIBEAN U.S.A.	21213	Broadway East	13
373	DUNKIN DONUTS	21212	Homeland	4
383	EASTSIDE SPORTS SOCIAL CLUB	21213	Broadway East	13
417	FIELDS OLD TRAIL	21212	Mid-Govans	4
475	GRAND CRU	21212	Chinquapin Park-Belvedere	4
545	RANDY'S BAR	21213	Broadway East	12
604	MURPHY'S NEIGHBORHOOD BAR & GRILL	21212	Mid-Govans	4

Cross tabs

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

Admit	Gender	Dept	Freq
Admitted:12	Male :12	A:4	Min. : 8
Rejected:12	Female:12	B:4	1st Qu.: 80
		C:4	Median :170
		D:4	Mean :189
		E:4	3rd Qu.:302
		F:4	Max. :512

Cross tabs

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

	Admit	
Gender	Admitted	Rejected
Male	1198	1493
Female	557	1278

Flat tables

```
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
```

```
, , replicate = 3
```


Flat tables

```
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

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

800040 bytes

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

0.8 Mb