Baltimore Restaurant Data Cleansing Project

Alice Zhou

13/01/2021

Let's first load data. The data source is from Baltimore open data source website. It is a data set about restaurants in Baltimore.The file is loaded from the URL address and exported as a csv file. I first check if the data directory has already been created. If not, I can create a brand new one. Recording today's date will help us update the data in the future.

```
setwd("C:/Users/alice/OneDrive/Desktop/Jobs/Data Cleaning")
if (!file.exists("data")) {
  dir.create("data")
}
fileUrl =
"https://opendata.baltimorecity.gov/egis/rest/services/Hosted/Red_Light_Camer
as/FeatureServer/0/query?outFields=*&where=1%3D1"
download.file(fileUrl, destfile = "./data/cameras.csv", method = "curl")
list.files("./data")
```

```
## [1] "cameras.csv"     "Restaurants.csv"
```

```
dateDownloaded = date()
dateDownloaded
```

```
## [1] "Wed Jan 13 18:11:48 2021"
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse
1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Let's take a look at the raw data.

```
setwd("C:/Users/alice/OneDrive/Desktop/Jobs/Data Cleaning")
rest = read.csv(file = 'data/restaurants.csv')
head(rest)
```

```
##        ï..X        Y fid gis_id srcid_t srcid_i              edit_date
ftype
## 1 -76.56222 39.33063   1   27_1      NA       0 2008/06/22 00:00:00+00
27
## 2 -76.58941 39.28473   2   27_2      NA       0 2008/06/22 00:00:00+00
27
## 3 -76.57606 39.28220   3   27_3      NA       0 2008/06/22 00:00:00+00
27
## 4 -76.63307 39.33710   4   27_4      NA       0 2008/06/22 00:00:00+00
27
## 5 -76.65577 39.28350   5   27_5      NA       0 2008/06/22 00:00:00+00
27
## 6 -76.59115 39.32081   6   27_6      NA       0 2008/06/22 00:00:00+00
27
##   subtype loc_type loc_meth street_tag prcl_pin             address
city
## 1      NA       ST   GC_300    9.3e+14 5936A004   4509  BELAIR ROAD
Baltimore
## 2      NA       ST   GC_300    9.3e+14  1830009      1919  FLEET ST
Baltimore
## 3      NA       ST   GC_300    9.3e+14  1878005     2844  HUDSON ST
Baltimore
## 4      NA       ST   GC_300    9.3e+14  3563006    3998  ROLAND AVE
Baltimore
## 5      NA       ST   GC_300    9.3e+14 2174A001 2481  frederick ave
Baltimore
## 6      NA       ST   GC_300    9.3e+14  4128031    2722  HARFORD RD
Baltimore
##   state zipcode x_coord  y_coord                  name alias1      nghbrhd
## 1    MD   21206 1436170 606298.6                   410     NA    Frankford
## 2    MD   21231 1428555 589545.9                  1919     NA  Fells Point
## 3    MD   21224 1432338 588642.1                 SAUTE     NA       Canton
## 4    MD   21211 1416120 608569.0    #1 CHINESE KITCHEN     NA      Hampden
## 5    MD   21223 1409774 589021.9 #1 chinese restaurant     NA     Millhill
## 6    MD   21218 1428002 602683.5              19TH HOLE     NA Clifton Park
##   cncldst    stfid_blk plcdst_no       plcdst               usng cntct_nme
## 1       2 2.451026e+14         4 NORTHEASTERN 18S UJ 65353 54630        NA
## 2       1 2.451020e+13         2 SOUTHEASTERN 18S UJ 62920 49577        NA
## 3       1 2.451010e+13         2 SOUTHEASTERN 18S UJ 64067 49276        NA
## 4      14 2.451013e+14         5     NORTHERN 18S UJ 59260 55457        NA
## 5       9 2.451020e+14         8 SOUTHWESTERN 18S UJ 57194 49544        NA
## 6      14 2.451080e+13         4 NORTHEASTERN 18S UJ 62840 53584        NA
##   cntct_phn cntct_dpt                                globalid url
## 1        NA        NA {B539CABF-622F-4C6E-9AE2-5C648EBBA530}  NA
## 2        NA        NA {319B2500-6082-4DD9-A866-66308527E3B2}  NA
## 3        NA        NA {57842273-D14B-4D28-A810-A645F018BED4}  NA
## 4        NA        NA {2701DA61-1457-496D-8743-A80949B80232}  NA
## 5        NA        NA {D51ACB6C-38C9-4C55-B221-FE47CE0D7C05}  NA
## 6        NA        NA {47F28D66-FE1E-4B2B-BF67-F952B94567B5}  NA
```

The key process of data cleaning is to study the data set, identifying any quirks, weird issues, missing values, or other problems to address before I can do downstream analysis. I use summary and str commands to get an overall summary of the data set. It displays some information about every single variable that we have.

```
str(rest)
```

```
## 'data.frame':    1327 obs. of  32 variables:
##  $ ï..X      : num  -76.6 -76.6 -76.6 -76.6 -76.7 ...
##  $ Y         : num  39.3 39.3 39.3 39.3 39.3 ...
##  $ fid       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ gis_id    : chr  "27_1" "27_2" "27_3" "27_4" ...
##  $ srcid_t   : logi  NA NA NA NA NA NA ...
##  $ srcid_i   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ edit_date : chr  "2008/06/22 00:00:00+00" "2008/06/22 00:00:00+00"
"2008/06/22 00:00:00+00" "2008/06/22 00:00:00+00" ...
##  $ ftype     : int  27 27 27 27 27 27 27 27 27 27 ...
##  $ subtype   : logi  NA NA NA NA NA NA ...
##  $ loc_type  : chr  "ST" "ST" "ST" "ST" ...
##  $ loc_meth  : chr  "GC_300" "GC_300" "GC_300" "GC_300" ...
##  $ street_tag: num  9.3e+14 9.3e+14 9.3e+14 9.3e+14 9.3e+14 ...
##  $ prcl_pin  : chr  "5936A004" "1830009" "1878005" "3563006" ...
##  $ address   : chr  "4509  BELAIR ROAD" "1919  FLEET ST" "2844  HUDSON ST"
"3998  ROLAND AVE" ...
##  $ city      : chr  "Baltimore" "Baltimore" "Baltimore" "Baltimore" ...
##  $ state     : chr  "MD" "MD" "MD" "MD" ...
##  $ zipcode   : chr  "21206" "21231" "21224" "21211" ...
##  $ x_coord   : num  1436170 1428555 1432338 1416120 1409774 ...
##  $ y_coord   : num  606299 589546 588642 608569 589022 ...
##  $ name      : chr  "410" "1919" "SAUTE" "#1 CHINESE KITCHEN" ...
##  $ alias1    : logi  NA NA NA NA NA NA ...
##  $ nghbrhd   : chr  "Frankford" "Fells Point" "Canton" "Hampden" ...
##  $ cncldst   : int  2 1 1 14 9 14 13 7 13 1 ...
##  $ stfid_blk : num  2.45e+14 2.45e+13 2.45e+13 2.45e+14 2.45e+14 ...
##  $ plcdst_no : int  4 2 2 5 8 4 2 5 2 2 ...
##  $ plcdst    : chr  "NORTHEASTERN" "SOUTHEASTERN" "SOUTHEASTERN"
"NORTHERN" ...
##  $ usng      : chr  "18S UJ 65353 54630" "18S UJ 62920 49577" "18S UJ
64067 49276" "18S UJ 59260 55457" ...
##  $ cntct_nme : logi  NA NA NA NA NA NA ...
##  $ cntct_phn : logi  NA NA NA NA NA NA ...
##  $ cntct_dpt : logi  NA NA NA NA NA NA ...
##  $ globalid  : chr  "{B539CABF-622F-4C6E-9AE2-5C648EBBA530}" "{319B2500-
6082-4DD9-A866-66308527E3B2}" "{57842273-D14B-4D28-A810-A645F018BED4}"
"{2701DA61-1457-496D-8743-A80949B80232}" ...
##  $ url       : logi  NA NA NA NA NA NA ...
```

Another way to look at the data is to focus on each variable at a time. Here, a zipcode table is created to see the number of restaurants under each zip code. Missing values will also be displayed if there are any by adding the second part to the command below.

```
table(rest$zipcode, useNA="ifany")
```

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

Here, I make a two-dimensional table with variables council districts and zip code. There are 37 restaurants in council district 1 in the 21202 zipcodes.

```
head(table(rest$cnclds, rest$zipcode))
```

```
##
##     21201 21202 21205 21206 21207 21208 21209 21210 21211 21212 21213
21214
##   1     0    37     0     0     0     0     0     0     0     0     2
0
##   2     0     0     3    27     0     0     0     0     0     0     0
0
##   3     0     0     0     0     0     0     0     0     0     0     2
17
##   4     0     0     0     0     0     0     0     0     0    27     0
0
##   5     0     0     0     0     3     0     6     0     0     0     0
0
##   6     0     0     0     0     0     0     1    19     0     0     0
0
##
##     21215 21216 21217 21218 21220 21222 21223 21224 21225 21226 21226-
21227
##   1     0     0     0     0     0     7     0   140     1     0      0
0
##   2     0     0     0     0     0     0     0    54     0     0      0
0
##   3     0     0     0     3     0     0     0     0     0     0      0
1
##   4     0     0     0     0     0     0     0     0     0     0      0
0
##   5    31     0     0     0     0     0     0     0     0     0      0
0
##   6    15     1     0     0     0     0     0     0     0     0      0
0
##
##     21229 21230 21231 21234 21237 21239 21251 21287
```

```
##   1     0     1   124     0     0     0     0     0
##   2     0     0     0     0     1     0     0     0
##   3     0     0     0     7     0     0     2     0
##   4     0     0     0     0     0     3     0     0
##   5     0     0     0     0     0     0     0     0
##   6     0     0     0     0     0     0     0     0
```

Let's take a look at the quantitative data.

```
quantile(rest$cncldst, na.rm=TRUE)

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

Now I am interested in all restaurants in the neighbourhood near me, Roland and Homeland.

```
rest$nearMe = rest$nghbrhd %in% c("Roland Park","Homeland")
table(rest$nearMe)

##
## FALSE  TRUE
##  1314    13
```

The next thing I want to do is to check for missing values of the council district variable. It is useful to use the is.na function to take the sum of the kind that is.na.

```
sum(is.na(rest$cncldst))

## [1] 0
```

There are 4 empty columns.

```
colSums(is.na(rest))

##       ï..X          Y        fid     gis_id    srcid_t    srcid_i
edit_date
##          0          0          0          0       1327          0
0
##      ftype    subtype   loc_type   loc_meth street_tag   prcl_pin
address
##          0       1327          0          0          0          0
0
##       city      state    zipcode    x_coord    y_coord       name
alias1
##          0          0          0          0          0          0
1327
##    nghbrhd    cncldst  stfid_blk  plcdst_no     plcdst       usng
cntct_nme
##          0          0          0          0          0          0
1327
```

```
##  cntct_phn  cntct_dpt   globalid        url     nearMe
##       1327       1327          0       1327          0
```

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

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

Are there any zip code either equal to 21212 or 21213? There are 59 of them in total.

```
table(rest$zipcode %in% c("21212","21213"))
```

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

Now I want to check if the zip codes and the council districts are wrong. It returns TRUE if a value is less than or equal to zero and returns false otherwise.

```
rest$zipWrong = ifelse(rest$zipcode <= 0, TRUE, FALSE)
table(rest$zipWrong, rest$zipcode < 0)
```

```
##
##         FALSE
##   FALSE  1327
```

```
rest$councilWrong = ifelse(rest$cncldst <= 0, TRUE, FALSE)
table(rest$councilWrong, rest$cncldst < 0)
```

```
##
##         FALSE
##   FALSE  1327
```

Eventually, I want only the restaurants with zip code 21212 and 21213 by creating a subset of the restaurant data with row condition specified.

```
rest.data = rest[rest$zipcode %in% c("21212","21213"),]
head(rest.data)
```

```
##          ï..X        Y fid gis_id srcid_t srcid_i              edit_date
ftype
## 29  -76.61164 39.28893  29  27_96      NA       0 2008/06/22 00:00:00+00
27
## 39  -76.59237 39.31225  39 27_106      NA       0 2008/06/22 00:00:00+00
27
## 92  -76.60952 39.36416  92  27_63      NA       0 2008/06/22 00:00:00+00
27
## 111 -76.58323 39.31329 111  27_82      NA       0 2008/06/22 00:00:00+00
27
## 187 -76.61105 39.36411 187 27_187      NA       0 2008/06/22 00:00:00+00
27
## 220 -76.60952 39.36416 220 27_220      NA       0 2008/06/22 00:00:00+00
27
```

```
##     subtype loc_type loc_meth street_tag prcl_pin             address
city
## 29       NA       ST   GC_300    9.3e+14  0662010    206 E REDWOOD ST
Baltimore
## 39       NA       ST   GC_300    9.3e+14  4156023    1801 E NORTH AVE
Baltimore
## 92       NA       ST   GC_300    9.3e+14  5134019 529 E BELVEDERE AVE
Baltimore
## 111      NA       ST   GC_300    9.3e+14  4177038     1932  BELAIR RD
Baltimore
## 187      NA       ST   GC_300    9.3e+14 5093B045 438 E BELVEDERE AVE
Baltimore
## 220      NA       ST   GC_300    9.3e+14  5134019 529 E BELVEDERE AVE
Baltimore
##     state zipcode x_coord  y_coord                       name alias1
## 29     MD   21212 1422255 591048.7          BAY ATLANTIC CLUB     NA
## 39     MD   21213 1427671 599567.3                BERMUDA BAR     NA
## 92     MD   21212 1422737 618452.8                  ATWATER'S     NA
## 111    MD   21213 1430256 599955.0 BALTIMORE ESTONIAN SOCIETY     NA
## 187    MD   21212 1422307 618430.4                   CAFE ZEN     NA
## 220    MD   21212 1422737 618452.8        CERIELLO FINE FOODS     NA
##                        nghbrhd cncldst    stfid_blk plcdst_no   plcdst
## 29                    Downtown      11 2.451040e+13         1  CENTRAL
## 39                Broadway East      12 2.451081e+13         3  EASTERN
## 92  Chinquapin Park-Belvedere       4 2.451027e+14         5 NORTHERN
## 111        South Clifton Park      12 2.451080e+13         3  EASTERN
## 187                   Rosebank       4 2.451027e+14         5 NORTHERN
## 220 Chinquapin Park-Belvedere       4 2.451027e+14         5 NORTHERN
##                     usng cntct_nme cntct_phn cntct_dpt
## 29  18S UJ 61011 50077        NA        NA        NA
## 39  18S UJ 62718 52637        NA        NA        NA
## 92  18S UJ 61342 58424        NA        NA        NA
## 111 18S UJ 63509 52737        NA        NA        NA
## 187 18S UJ 61211 58420        NA        NA        NA
## 220 18S UJ 61342 58424        NA        NA        NA
##                                     globalid url nearMe zipWrong
councilWrong
## 29  {81105212-0BFB-4A8C-8FA9-670888C91817}  NA  FALSE    FALSE
FALSE
## 39  {C4D8D749-5E9D-4454-ADB9-0F14E07336EE}  NA  FALSE    FALSE
FALSE
## 92  {EF3876DC-4041-4A93-A16A-5B50DD3599B7}  NA  FALSE    FALSE
FALSE
## 111 {A41B4B47-062F-45CD-A089-FE4234620971}  NA  FALSE    FALSE
FALSE
## 187 {697308FD-325A-4492-A1EB-241156A3D2D2}  NA  FALSE    FALSE
FALSE
## 220 {FE9E86BC-9F28-4F02-9D78-6DD297286124}  NA  FALSE    FALSE
FALSE
```

I can get rid of the columns with missing values and only display the columns that are meaningful to the analysis.

```
rest.data = rest.data %>% select(14,15,17,20,22,23,25,26)
head(rest.data)
```

```
##                 address      city zipcode                       name
## 29     206 E REDWOOD ST Baltimore   21212          BAY ATLANTIC CLUB
## 39     1801 E NORTH AVE Baltimore   21213                BERMUDA BAR
## 92  529 E BELVEDERE AVE Baltimore   21212                  ATWATER'S
## 111     1932  BELAIR RD Baltimore   21213 BALTIMORE ESTONIAN SOCIETY
## 187 438 E BELVEDERE AVE Baltimore   21212                   CAFE ZEN
## 220 529 E BELVEDERE AVE Baltimore   21212        CERIELLO FINE FOODS
##                        nghbrhd cncldst plcdst_no   plcdst
## 29                    Downtown      11         1  CENTRAL
## 39                Broadway East      12         3  EASTERN
## 92  Chinquapin Park-Belvedere       4         5 NORTHERN
## 111        South Clifton Park      12         3  EASTERN
## 187                   Rosebank       4         5 NORTHERN
## 220 Chinquapin Park-Belvedere       4         5 NORTHERN
```

Now I order the data by the restaurant name alphabetically.Finally we have a clean data set of restaurants with zip code 21212 and 21213.

```
head(rest.data[order(rest.data$name),])
```

```
##                 address      city zipcode                       name
## 92  529 E BELVEDERE AVE Baltimore   21212                  ATWATER'S
## 111     1932  BELAIR RD Baltimore   21213 BALTIMORE ESTONIAN SOCIETY
## 29     206 E REDWOOD ST Baltimore   21212          BAY ATLANTIC CLUB
## 39     1801 E NORTH AVE Baltimore   21213                BERMUDA BAR
## 187 438 E BELVEDERE AVE Baltimore   21212                   CAFE ZEN
## 220 529 E BELVEDERE AVE Baltimore   21212        CERIELLO FINE FOODS
##                        nghbrhd cncldst plcdst_no   plcdst
## 92  Chinquapin Park-Belvedere       4         5 NORTHERN
## 111        South Clifton Park      12         3  EASTERN
## 29                    Downtown      11         1  CENTRAL
## 39                Broadway East      12         3  EASTERN
## 187                   Rosebank       4         5 NORTHERN
## 220 Chinquapin Park-Belvedere       4         5 NORTHERN
```

The last thing I want to know is the size of my data set. I use the object size command to see how many bytes that data set is. The newly processed data set size is a lot smaller than the raw data set size.

```
object.size(rest.data)
```

```
## 15240 bytes
```

```
object.size(rest)
```

```
## 848896 bytes
```