Maryland schools star ratings map

By Christine Zhang (mailto:czhang@baltsun.com)

This notebook appends latitude and longitude coordinates for schools in Maryland for mapping purposes.

The map can be found embedded in Baltimore Sun stories here (https://www.baltimoresun.com/news/maryland/education/k-12/bs-md-star-rating-release-20181203-story.html) and here (https://www.baltimoresun.com/news/maryland/education/k-12/bs-md-star-ratings-key-takeaways-20181204-story.html).

Geographical information for schools comes from National Center for Education Statistics 2016-17 Education Demographic and Geographic Estimates (EDGE) (https://nces.ed.gov/programs/edge/Geographic/SchoolLocations).

How we did it

Import R data analysis libraries and read in star ratings data

In [1]:
```
suppressMessages(library('tidyverse'))
suppressMessages(library('stringr'))
suppressMessages(library('janitor'))
```

Read in the scores data.

In [2]:
```
scores <- suppressMessages(read_csv('input/accountability_schools_download_file.csv', na = 'na') %>% clean_names())
```

Schools in the star ratings data are uniquely identified by a combination of the `lea_number` and `school_number`.

In [3]:
```
glimpse(scores)
```

```
Observations: 1,319
Variables: 10
$ number_academic_year        <int> 2018, 2018, 2018, 2018, 2018, 2018, 201...
$ lea_number                  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
$ lea_name                    <chr> "Allegany", "Allegany", "Allegany", "Al...
$ school_number               <int> 301, 401, 402, 405, 406, 502, 504, 601,...
$ school_name                 <chr> "Flintstone Elementary", "South Penn El...
$ star_rating                 <int> 4, 4, 4, 3, 3, 5, 3, 4, 5, 4, 5, 4, 4, ...
$ total_earned_points_percent <int> 64, 65, 69, 59, 56, 79, 58, 60, 78, 64,...
$ percentile_rank_elementary  <int> 52, 52, 67, NA, NA, 91, NA, NA, 88, NA,...
$ percentile_rank_middle      <int> NA, NA, NA, NA, 49, NA, 55, NA, NA, NA,...
$ percentile_rank_high        <int> NA, NA, NA, 45, NA, NA, NA, 48, NA, 57,...
```

Read in the EDGE data, which provides coordinates for schools nationwide

In [4]:
```
edge <- read.csv('input/EDGE_GEOCODE_PUBLICSCH_1617.csv', stringsAsFactors = F,
                 colClasses = c('NCESSCH' = 'character')) %>% clean_names()
```

Schools in the EDGE data are uniquely idenfified by the 12-digit `ncessch` number (it's irrelevant for MD schools, but we specify `colClasses = c('NCESSCH' = 'character')` so that `R` will not drop the leading zero.

In [5]:
```
glimpse(edge)
```

```
Observations: 102,173
Variables: 24
$ ncessch  <chr> "010000200277", "010000201667", "010000201670", "010000201...
$ name     <chr> "Sequoyah Sch - Chalkville Campus", "Camps", "Det Ctr", "W...
$ opstfips <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
$ street   <chr> "1000 Industrial School Road", "1601 County Rd. 57", "2109...
$ city     <chr> "Birmingham", "Prattville", "Thomasville", "Mount Meigs", ...
$ state    <chr> "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL"...
$ zip      <int> 35220, 36067, 36784, 36057, 35206, 36057, 35950, 35950, 35...
$ stfip    <chr> "01", "01", "01", "01", "01", "01", "01", "01", "01", "01"...
$ cnty     <chr> "01073", "01001", "01025", "01101", "01073", "01101", "010...
$ nmcnty   <chr> "Jefferson County", "Autauga County", "Clarke County", "Mo...
$ locale   <chr> "21", "41", "41", "41", "12", "41", "32", "32", "32", "32"...
$ lat      <dbl> 33.67366, 32.51917, 31.93779, 32.37571, 33.58671, 32.37571...
$ lon      <dbl> -86.62875, -86.53275, -87.75016, -86.08321, -86.71058, -86...
$ cbsa     <chr> "13820", "33860", "N", "33860", "13820", "33860", "10700",...
$ nmcbsa   <chr> "Birmingham-Hoover, AL", "Montgomery, AL", "N", "Montgomer...
$ cbsatype <int> 1, 1, 0, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2...
$ csa      <chr> "142", "N", "N", "N", "142", "N", "290", "290", "290", "29...
$ nmcsa    <chr> "Birmingham-Hoover-Talladega, AL", "N", "N", "N", "Birming...
$ necta    <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N"...
$ nmnecta  <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N"...
$ cd       <chr> "0106", "0102", "0107", "0103", "0107", "0103", "0104", "0...
$ sldl     <chr> "01044", "01042", "01068", "01075", "01058", "01075", "010...
$ sldu     <chr> "01020", "01030", "01024", "01025", "01020", "01025", "010...
$ survyear <int> 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016...
```

Read in the school directory data

We can't directly match up schools in the star ratings data with schools in the EDGE data because they are have different identifiers. Here we read in the Maryland school directory from the Maryland State Department of Education website (http://reportcard.msde.maryland.gov/). This file provides the a way to link the two datasets.

In [14]:
```
directory <- suppressMessages(read.csv('input/School_Directory_2018.csv',
                                       colClasses = c('LEA.Number' = 'character',
                                                      'School.Number' = 'character
',
                                                      'NCES.Number' = 'character'))
%>% clean_names())
```

Schools are identified by `lea_number` and `school_number`.

In [15]:
```
glimpse(directory)
```

```
Observations: 1,429
Variables: 13
$ academic_year <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018,...
$ lea_number    <chr> "01", "01", "01", "01", "01", "01", "01", "01", "01",...
$ lea_name      <fct> Allegany, Allegany, Allegany, Allegany, Allegany, All...
$ school_number <chr> "0301", "0401", "0402", "0405", "0406", "0502", "0504...
$ school_name   <fct> Flintstone Elementary, South Penn Elementary, John Hu...
$ school_type   <fct> E, E, E, H, M, E, M, H, , E, H, E, E, E, M, E, E, M, ...
$ address       <fct> 22000 National Pike Ne, 500 E 2nd St, 120 Mary St, 50...
$ city          <fct> Flintstone, Cumberland, Cumberland, Cumberland, Cumbe...
$ state         <fct> MD, MD, MD, MD, MD, MD, MD, MD, MD, MD, MD, MD, MD, M...
$ zip           <int> 21530, 215024249, 215027341, 215023856, 215023855, 21...
$ phone         <dbl> 3014782434, 3017771755, 3017248842, 3017772570, 30177...
$ nces_number   <chr> "240003000014", "240003001359", "240003000019", "2400...
$ create_date   <int> 20180820, 20180820, 20180820, 20180820, 20180820, 201...
```

Merge `scores` with `directory` to get the NCES id for each school

We can merge the `scores` and the `directory` dataframes on the `lea_number` and `school_number` columns. However, we first need to add a leading zero to `lea_number` and `school_number` in the `scores` dataframe. We can do this using `str_pad()` .

In [16]:
```
scores$school_number <- str_pad(scores$school_number, 4, pad = '0')
scores$lea_number <- str_pad(scores$lea_number, 2, pad = '0')
```

We will call the merged dataframe `scores.nces` .

In [17]:
```
scores.nces <- merge(scores, directory %>% select(-lea_name, -school_name),
                     by = c('lea_number', 'school_number'), all.x = T)
```

Merge `scores.nces` with `edge` to get the geographical coordinates for each school

We can merge the `scores.nces` and the `edge` dataframes on the `nces_number` (from `scores.nces`) and `ncessch` (from `edge`). This is the 12-digit NCES id for each school. We will call the merged dataframe `scores.geo` .

In [22]:
```
scores.geo <- merge(scores.nces, edge,
                    by.x = 'nces_number',
                    by.y = 'ncessch', all.x = T,
                    suffixes = c('_msde', '_nces'))
```

Note: there are three schools that do no have coordinates provided by EDGE.

In [24]:
```
scores.geo %>% filter(is.na(lat))
```

nces_number	lea_number	school_number	number_academic_year	lea_name	school_name	star_rating	tota
240006001744	02	6123	2018	Anne Arundel	Monarch Academy Annapolis ES	2	
240048001741	15	0835	2018	Montgomery	Silver Creek Middle	4	
240057001743	19	0107	2018	Somerset	Greenwood Elementary School	3	

We can add in the coordinates for these schools manually.

In [25]:
```
added <- suppressMessages(read_csv('input/addresses_add.csv'))
```

In [26]:
```
scores.geo.added <- merge(scores.geo, added, by = c('lea_number', 'school_number'),
all.x = T)
```

In [27]:
```
scores.geo.added <- scores.geo.added %>% mutate(lat = ifelse(is.na(lat.x), lat.y, l
at.x),
                                lon = ifelse(is.na(lon.x), lon.y, lon.x),
                                address = ifelse(is.na(address.x), as.character(addres
s.y), as.character(address.x)),
                                city_msde = ifelse(is.na(city_msde.x), as.character(cit
y_msde.y), as.character(city_msde.x)))
```

In [28]:
```
scores.geo.added <- scores.geo.added %>% select(lea_number,
                                                lea_name,
                                                school_number,
                                                school_name = school_name.x,
                                                nces_number,
                                                number_academic_year,
                                                star_rating,
                                                total_earned_points_percent,
                                                percentile_rank_elementary,
                                                percentile_rank_middle,
                                                percentile_rank_high,
                                                address,
                                                city = city_msde,
                                                lat,
                                                lon)
```

In [29]:
```
head(scores.geo.added)
```

lea_number	lea_name	school_number	school_name	nces_number	number_academic_year	star_rating	total_
01	Allegany	0301	Flintstone Elementary	240003000014	2018	4	
01	Allegany	0401	South Penn Elementary	240003001359	2018	4	
01	Allegany	0402	John Humbird Elementary	240003000019	2018	4	
01	Allegany	0405	Fort Hill High	240003000015	2018	3	
01	Allegany	0406	Washington Middle	240003000031	2018	3	
01	Allegany	0502	Northeast Elementary	240003000024	2018	5	

Write to `output/` folder

In [30]:
```
write_csv(scores.geo.added, 'output/scores_clean.csv')
```