

Tutorial 5: Maps I

Leah Brooks

February 17, 2020

This is our first of two tutorials on maps, concentrating on the R package `sf`. This tutorial introduces digital maps. It then gives examples of how to present them, followed by some examples of the type of spatial analysis you can do with this package. In our second mapping class, we concentrate on choropleth maps.

This tutorial begins by showing you how to make a simple map, and it ends with relatively sophisticated programming showing you how to count points within polygons. Specifically, here are the steps you'll take

- A) Load the package you'll need
- B) Draw map of DC wards, label the wards, and add map of DC school administration buildings
- C) Apply these techniques with bigger point data
- D) Count points in polygons

While making this tutorial, I consulted three online tutorials that I would recommend if you want more in-depth coverage of specific map issues. See

- The University of Chicago's [Computing for the Social Sciences](#)
- The [authors of the the package](#), in-depth, but sometimes difficult to follow
- A [random guy](#)

A. Load packages

This week we are adding a new package to our repertoire: `sf`. This package is a quantum leap forward for mapping in R. It is fast and, relative to what was previously available, easy to use (you may not believe this after today's tutorial, but it is true). This package is also designed to work with `ggplot`, so today's lesson shows you how these packages integrate.

As a word of warning, `sf` is very new. As such, the online help is not as extensive as for `ggplot`, and every so often you run into odd errors, or things that you think the package should do but does not. Despite all this, we are learning it because I think it's the best for mapping and spatial analysis in R: comprehensive and fast.

All `sf` commands begin with `st_`, as you'll see in this tutorial. We introduce only a fraction of what `sf` can do in this tutorial, so look at the online help or discuss with me if you have more questions.

Begin by installing the `sf` package, using the `dependencies = TRUE` option. This option automatically installs any packages that `sf` needs that you don't have. Recall that you install packages once (and for all) by typing in the console window. Everything else in this tutorial should be written in a R script.

```
install.packages("sf", dependencies = TRUE)
```

We also load `ggplot2` and `dplyr` – packages you've already installed. Here I use the `require()` command. This command checks if you have the package. If you do, it loads the package. If it doesn't it downloads the package and then loads the package.

Now load the packages we'll need for today.

```
require(sf)
```

```
## Loading required package: sf  
## Linking to GEOS 3.6.1, GDAL 2.2.3, PROJ 4.9.3
```

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
require(dplyr)
```

```
## Loading required package: dplyr  
##  
## Attaching package: 'dplyr'  
## The following objects are masked from 'package:stats':  
##  
##   filter, lag  
## The following objects are masked from 'package:base':  
##  
##   intersect, setdiff, setequal, union
```

B. Load, explore and plot two shapefiles

In this section we load a shapefile of DC wards, label them, and plot one additional feature on top of the ward maps.

B.1. Load

We begin by loading a shapefile of wards in DC. There are 8 wards in DC (electoral districts for councilmembers – there are also 5 at-large councilmembers), so we expect this file to have 8 polygons. You can download this file from DC’s Open Data site [here](#). From the download menu, choose “shapefile” and then save in a location you’ll remember.

To load a shapefile into R with `sf`, you use the `st_read()` command. You tell R where the file is, and what type of file it is by the extension.

The `st_read()` command is specifically designed to work with shapefiles, which are files that contain spatial data. For more details on such files, see the lecture notes.

Begin by loading the data, as below.

```
dc.wards <- st_read("H:/pppa_data_viz/2019/tutorial_data/lecture05/Ward_from_2012/Ward_from_2012.shp")  
  
## Reading layer `Ward_from_2012' from data source `H:\pppa_data_viz\2019\tutorial_data\lecture05\Ward_  
## Simple feature collection with 8 features and 82 fields  
## geometry type: POLYGON  
## dimension: XY  
## bbox: xmin: -77.1198 ymin: 38.79164 xmax: -76.90915 ymax: 38.99597  
## epsg (SRID): 4326  
## proj4string: +proj=longlat +datum=WGS84 +no_defs
```

Note that `sf` tells you a little about this file immediately upon loading. The “8 features” to which R refers are the 8 wards, so features are the geographic units of analysis. We also learn that the “geometry type” is

this file is polygons (not lines or points). R also tells us maximum and minimum latitude and longitude of the map and the projection number (more on this later).

B.2. Explore

One of the useful aspects of simple features (the name of features loaded via this package) is that you can use all your usual dataframe commands on them.

First try

```
head(dc.wards)
```

```
## Simple feature collection with 6 features and 82 fields
## geometry type: POLYGON
## dimension: XY
## bbox: xmin: -77.08172 ymin: 38.79164 xmax: -76.90915 ymax: 38.9573
## epsg (SRID): 4326
## proj4string: +proj=longlat +datum=WGS84 +no_defs
## OBJECTID WARD NAME REP_NAME
## 1 1 8 Ward 8 Trayon White, Sr.
## 2 2 6 Ward 6 Charles Allen
## 3 3 7 Ward 7 Vincent Gray
## 4 4 2 Ward 2 Jack Evans
## 5 5 1 Ward 1 Brianne Nadeau
## 6 6 5 Ward 5 Kenyan McDuffie
## WEB_URL REP_PHONE
## 1 http://dccouncil.us/council/trayon-white-sr (202) 724-8045
## 2 http://dccouncil.us/council/charles-allen (202) 724-8072
## 3 http://dccouncil.us/council/vincent-gray (202) 724-8068
## 4 http://dccouncil.us/council/jack-evans (202) 724-8058
## 5 http://dccouncil.us/council/brianne-nadeau (202) 724-8181
## 6 http://dccouncil.us/council/kenyan-mcduffie (202) 724-8028
## REP_EMAIL REP_OFFICE WARD_ID
## 1 twhite@dccouncil.us 1350 Pennsylvania Ave, Suite 400, NW 20004 8
## 2 callen@dccouncil.us 1350 Pennsylvania Ave, Suite 406, NW 20004 6
## 3 vgray@dccouncil.us 1350 Pennsylvania Ave, Suite 406, NW 20004 7
## 4 jevans@dccouncil.us 1350 Pennsylvania Ave, Suite 106, NW 20004 2
## 5 bnadeau@dccouncil.us 1350 Pennsylvania Ave, Suite 102, NW 20004 1
## 6 kmcduffie@dccouncil.us 1350 Pennsylvania Ave, Suite 506, NW 20004 5
## LABEL AREASQMI Shape_Leng Shape_Area POP_2000 POP_2010 POP_2011_2 POP_BLACK
## 1 Ward 8 11.937871 28714.07 30965852 74049 73662 81133 75259
## 2 Ward 6 6.221045 24157.98 16064917 70867 76238 84290 29909
## 3 Ward 7 8.809914 22345.23 22818183 69987 71748 73290 69005
## 4 Ward 2 8.684517 29545.80 22492798 63455 76645 77645 6817
## 5 Ward 1 2.535896 12925.38 6567941 71747 74462 82859 25110
## 6 Ward 5 10.390304 22893.40 26910761 71440 74308 82049 57733
## POP_NATIVE POP_ASIAN POP_HAWAII POP_OTHER TWO_OR_MOR NOT_HISPAN HISPANIC_0
## 1 110 310 12 711 872 79843 1290
## 2 295 3573 40 1233 2529 79000 5290
## 3 219 225 17 1211 908 70987 2303
## 4 213 7640 30 2496 2875 69529 8116
## 5 300 3509 111 6259 2596 65755 17104
## 6 335 1622 9 3758 1915 75058 6991
## POP_MALE POP_FEMALE AGE_0_5 AGE_5_9 AGE_10_14 AGE_15_17 AGE_18_19 AGE_20
```

## 1	35573	45560	7879	7061	5963	3596	2945	1800
## 2	40411	43879	4779	2747	2235	1088	1370	794
## 3	33916	39374	5230	4485	4333	2944	2162	1347
## 4	39214	38431	2173	1110	571	486	6200	3363
## 5	41368	41491	4733	2644	1934	1223	2399	1613
## 6	38692	43357	5778	3972	2654	2249	2817	1533
##	AGE_21	AGE_22_24	AGE_25_29	AGE_30_34	AGE_35_39	AGE_40_44	AGE_45_49	AGE_50_54
## 1	1748	4070	6306	5951	4617	4873	4429	4978
## 2	724	4596	13427	12512	8052	5474	4524	4350
## 3	1107	3170	5036	5083	4154	5166	4794	5715
## 4	3485	6463	12614	10475	6491	4175	3445	3413
## 5	1642	5487	15600	12286	7378	5099	4784	3963
## 6	1531	3882	8514	7439	5996	5228	4828	4905
##	AGE_55_59	AGE_60_61	AGE_65_66	AGE_67_69	AGE_70_74	AGE_75_79	AGE_80_84	
## 1	5001	1578	1011	1211	1904	1026	634	
## 2	4724	1830	1518	1596	2167	1522	816	
## 3	5272	1716	1254	1680	2370	1719	1290	
## 4	3248	1447	1214	1325	1763	764	724	
## 5	2618	1420	802	1208	1778	954	505	
## 6	5274	1736	1108	1640	2361	2079	1947	
##	AGE_85_PLU	MEDIAN_AGE	UNEMPLOYME	TOTAL_HH	FAMILY_HH	PCT_FAMILY		
## 1	432	29.3	22.9	29470	17747	60.2205632846963		
## 2	958	33.9	6.3	40100	15110	37.6807980049875		
## 3	1142	37	19.1	29266	15574	53.2153352012574		
## 4	863	30.9	3.7	38870	9071	23.3367635708773		
## 5	738	31.3	6.6	34907	12253	35.1018420374137		
## 6	2135	35.4	14.1	31656	14893	47.0463735152894		
##	NONFAMILY_	PCT_NONFAM	PCT_BELOW_	PCT_BELO_1	PCT_BELO_2	PCT_BELO_3		
## 1	11723	39.7794367153037		37.7	35.3	10.8	38.8	
## 2	24990	62.3192019950125		12.5	9.6	4.5	25.7	
## 3	13692	46.7846647987426		27.2	23.6	8.9	28	
## 4	29799	76.6632364291227		13.4	4.7	10.1	33.2	
## 5	22654	64.8981579625863		13.5	11	7.7	23.3	
## 6	16763	52.9536264847106		19	13.5	12.5	20.4	
##	PCT_BELO_4	PCT_BELO_5	PCT_BELO_6	PCT_BELO_7	PCT_BELO_8	POP_25_PLU	POP_25_P_1	
## 1	54.1	21.6	0	51.2	36.4	<NA>	1858	
## 2	13.9	11.6	0	10.7	6.2	<NA>	1785	
## 3	12.8	5.3	0	19.7	20	<NA>	2259	
## 4	26	23.5	0	12.6	10.3	<NA>	1743	
## 5	8.7	6.9	0	26.8	11.5	<NA>	4324	
## 6	19.1	17.7	0	27.9	10.4	<NA>	2796	
##	POP_25_P_2	MARRIED_CO	MALE_HH_NO	FEMALE_HH_	MEDIAN_HH_	PER_CAPITA	PCT_BELO_9	
## 1	2614	<NA>	1886	11653	30910	17596	36.3	
## 2	25882	<NA>	944	4157	94343	58354	6.3	
## 3	3309	<NA>	1682	9499	39165	22917	22.9	
## 4	28509	<NA>	458	754	100388	72388	13.5	
## 5	22488	<NA>	1634	3133	82159	47982	17.9	
## 6	11168	<NA>	1567	6726	57554	32449	21.6	
##	PCT_BELO10	NO_DIPLOMA	DIPLOMA_25	NO_DEGREE_	ASSOC_DEGR	BACH_DEGRE	MED_VAL_00	
## 1	10.8	5958	18736	10975	2149	3781	229900	
## 2	4.5	3128	7079	6643	1852	19588	573200	
## 3	8.9	6002	18683	10800	2569	4890	238900	
## 4	10.1	988	2381	2980	940	16253	623500	
## 5	7.7	3224	6984	5293	1258	17613	542100	

```
## 6      12.5      5094      13788      11034      2196      11557      379800
## Shape_Le_1 Shape_Ar_1 geometry
## 1  28714.07  30965852 POLYGON ((-76.97229 38.8728...
## 2  24157.98  16064917 POLYGON ((-77.0179 38.9141,...
## 3  22345.23  22818183 POLYGON ((-76.94186 38.9185...
## 4  29545.80  22492798 POLYGON ((-77.04946 38.9199...
## 5  12925.38   6567941 POLYGON ((-77.03523 38.9374...
## 6  22893.40  26910761 POLYGON ((-76.99144 38.9573...
```

This shows some summary information on the geography part of the file – the type of features, the coordinate system – and then the first six rows of the data portion of the file.

We can also use `str()` to get a description of all the variables:

```
str(dc.wards)
```

```
## Classes 'sf' and 'data.frame':  8 obs. of  83 variables:
## $ OBJECTID : num  1 2 3 4 5 6 7 8
## $ WARD      : num  8 6 7 2 1 5 3 4
## $ NAME      : Factor w/ 8 levels "Ward 1","Ward 2",...: 8 6 7 2 1 5 3 4
## $ REP_NAME  : Factor w/ 8 levels "Brandon T. Todd",...: 7 3 8 4 2 5 6 1
## $ WEB_URL   : Factor w/ 8 levels "http://dccouncil.us/council/brandon-todd",...: 7 3 8 4 2 5 6 1
## $ REP_PHONE : Factor w/ 8 levels "(202) 724-8028",...: 2 7 6 4 8 1 5 3
## $ REP_EMAIL : Factor w/ 8 levels "bnadeau@dccouncil.us",...: 7 3 8 4 1 5 6 2
## $ REP_OFFICE: Factor w/ 7 levels "1350 Pennsylvania Ave, Suite 102, NW 20004",...: 5 6 6 3 1 7 4 2
## $ WARD_ID   : Factor w/ 8 levels "1","2","3","4",...: 8 6 7 2 1 5 3 4
## $ LABEL     : Factor w/ 8 levels "Ward 1","Ward 2",...: 8 6 7 2 1 5 3 4
## $ AREASQMI  : num  11.94 6.22 8.81 8.68 2.54 ...
## $ Shape_Leng: num  28714 24158 22345 29546 12925 ...
## $ Shape_Area: num  30965852 16064917 22818183 22492798 6567941 ...
## $ POP_2000  : num  74049 70867 69987 63455 71747 ...
## $ POP_2010  : num  73662 76238 71748 76645 74462 ...
## $ POP_2011_2: num  81133 84290 73290 77645 82859 ...
## $ POP_BLACK : Factor w/ 8 levels "25110","29909",...: 8 2 7 6 1 5 4 3
## $ POP_NATIVE: Factor w/ 8 levels "110","163","213",...: 1 5 4 3 6 7 2 8
## $ POP_ASIAN : Factor w/ 8 levels "1622","1755",...: 4 6 3 8 5 1 7 2
## $ POP_HAWAII: Factor w/ 8 levels "0","111","12",...: 3 7 4 5 2 8 1 6
## $ POP_OTHER_: Factor w/ 8 levels "1211","1233",...: 7 2 1 4 6 5 3 8
## $ TWO_OR_MOR: Factor w/ 8 levels "1915","2287",...: 7 3 8 5 4 1 6 2
## $ NOT_HISPAN: Factor w/ 8 levels "65755","66565",...: 8 7 4 3 1 6 5 2
## $ HISPANIC_0: Factor w/ 8 levels "1290","16501",...: 1 5 4 7 3 6 8 2
## $ POP_MALE  : Factor w/ 8 levels "33916","35573",...: 2 7 1 5 8 4 3 6
## $ POP_FEMALE: Factor w/ 8 levels "38431","39374",...: 7 6 2 1 3 5 8 4
## $ AGE_0_5   : Factor w/ 8 levels "2173","4259",...: 8 4 5 1 3 7 2 6
## $ AGE_5_9   : Factor w/ 8 levels "1110","2644",...: 8 3 7 1 2 4 5 6
## $ AGE_10_14 : Factor w/ 8 levels "1934","2235",...: 8 2 6 7 1 3 4 5
## $ AGE_15_17 : Factor w/ 8 levels "1088","1223",...: 7 1 6 8 2 4 3 5
## $ AGE_18_19 : Factor w/ 8 levels "1370","1379",...: 6 1 3 8 4 5 7 2
## $ AGE_20    : Factor w/ 8 levels "1347","1497",...: 5 8 1 6 4 3 2 7
## $ AGE_21    : Factor w/ 8 levels "1107","1499",...: 5 8 1 6 4 3 2 7
## $ AGE_22_24 : Factor w/ 8 levels "2893","3170",...: 5 6 2 8 7 4 3 1
## $ AGE_25_29 : Factor w/ 8 levels "12614","13427",...: 5 2 4 1 3 7 8 6
## $ AGE_30_34 : Factor w/ 8 levels "10475","12286",...: 5 3 4 1 2 8 7 6
## $ AGE_35_39 : Factor w/ 8 levels "4154","4617",...: 2 8 1 6 7 4 5 3
## $ AGE_40_44 : Factor w/ 8 levels "4175","4873",...: 2 6 4 1 3 5 7 8
## $ AGE_45_49 : Factor w/ 8 levels "3445","4429",...: 2 3 5 1 4 6 7 8
```

```

## $ AGE_50_54 : Factor w/ 8 levels "3413","3963",...: 5 3 7 1 2 4 6 8
## $ AGE_55_59 : Factor w/ 8 levels "2618","3248",...: 5 4 6 2 1 7 3 8
## $ AGE_60_61 : Factor w/ 8 levels "1420","1447",...: 3 6 4 2 1 5 7 8
## $ AGE_65_66 : Factor w/ 8 levels "1011","1108",...: 1 6 4 3 8 2 7 5
## $ AGE_67_69 : Factor w/ 8 levels "1208","1211",...: 2 4 6 3 1 5 8 7
## $ AGE_70_74 : Factor w/ 8 levels "1763","1778",...: 3 4 7 1 2 6 8 5
## $ AGE_75_79 : Factor w/ 8 levels "1026","1522",...: 1 2 3 7 8 5 6 4
## $ AGE_80_84 : Factor w/ 8 levels "1290","1349",...: 6 8 1 7 5 4 2 3
## $ AGE_85_PLU: Factor w/ 8 levels "1142","1837",...: 5 8 1 7 6 3 2 4
## $ MEDIAN_AGE: Factor w/ 7 levels "29.3","30.9",...: 1 4 6 2 3 5 6 7
## $ UNEMPLOYME: Factor w/ 7 levels "14.1","19.1",...: 3 5 2 4 6 1 4 7
## $ TOTAL_HH : Factor w/ 8 levels "29266","29470",...: 2 8 1 7 5 4 6 3
## $ FAMILY_HH : Factor w/ 8 levels "12253","14893",...: 7 3 4 8 1 2 5 6
## $ PCT_FAMILY: Factor w/ 8 levels "23.3367635708773",...: 8 3 6 1 2 5 4 7
## $ NONFAMILY_: Factor w/ 8 levels "11723","12840",...: 1 7 3 8 6 4 5 2
## $ PCT_NONFAM: Factor w/ 8 levels "39.7794367153037",...: 1 6 3 8 7 4 5 2
## $ PCT_BELOW_: Factor w/ 8 levels "11.9","12.5",...: 7 2 6 3 4 5 8 1
## $ PCT_BELO_1: Factor w/ 8 levels "1.9","11","13.5",...: 5 8 4 6 2 3 1 7
## $ PCT_BELO_2: Factor w/ 8 levels "10.1","10.8",...: 2 4 8 1 7 3 6 5
## $ PCT_BELO_3: Factor w/ 8 levels "14.6","20.4",...: 8 5 6 7 4 2 3 1
## $ PCT_BELO_4: Factor w/ 8 levels "0","12.8","13.9",...: 6 3 2 5 8 4 1 7
## $ PCT_BELO_5: Factor w/ 8 levels "10.3","11.6",...: 5 2 7 6 8 4 3 1
## $ PCT_BELO_6: Factor w/ 2 levels "0","29": 1 1 1 1 1 1 1 2
## $ PCT_BELO_7: Factor w/ 8 levels "10.7","12.6",...: 8 1 4 2 5 7 6 3
## $ PCT_BELO_8: Factor w/ 8 levels "10.3","10.4",...: 6 7 5 1 4 2 8 3
## $ POP_25_PLU: Factor w/ 0 levels: NA NA NA NA NA NA NA NA
## $ POP_25_P_1: Factor w/ 8 levels "1743","1785",...: 3 2 4 1 7 5 8 6
## $ POP_25_P_2: Factor w/ 8 levels "11168","15399",...: 5 4 8 6 3 1 7 2
## $ MARRIED_CO: Factor w/ 0 levels: NA NA NA NA NA NA NA NA
## $ MALE_HH_NO: Factor w/ 8 levels "1567","1591",...: 5 8 4 6 3 1 7 2
## $ FEMALE_HH_: Factor w/ 8 levels "11653","1511",...: 1 4 8 7 3 6 2 5
## $ MEDIAN_HH_: Factor w/ 8 levels "100388","112873",...: 3 8 4 1 7 5 2 6
## $ PER_CAPITA: Factor w/ 8 levels "17596","22917",...: 1 6 2 7 5 3 8 4
## $ PCT_BELO_9: Factor w/ 8 levels "13.5","13.7",...: 7 8 6 1 4 5 2 3
## $ PCT_BELO10: Factor w/ 8 levels "10.1","10.8",...: 2 4 8 1 7 3 6 5
## $ NO_DIPLOMA: Factor w/ 8 levels "3128","3224",...: 5 1 6 8 2 4 7 3
## $ DIPLOMA_25: Factor w/ 8 levels "11907","13788",...: 4 8 3 6 7 2 5 1
## $ NO_DEGREE_: Factor w/ 8 levels "10500","10800",...: 3 8 2 5 7 4 6 1
## $ ASSOC_DEGR: Factor w/ 8 levels "1003","1258",...: 5 3 7 8 2 6 1 4
## $ BACH_DEGRE: Factor w/ 8 levels "11557","13032",...: 7 6 8 3 4 1 5 2
## $ MED_VAL_00: Factor w/ 8 levels "229900","238900",...: 1 6 2 7 5 3 8 4
## $ Shape_Le_1: num 28714 24158 22345 29546 12925 ...
## $ Shape_Ar_1: num 30965852 16064917 22818183 22492798 6567941 ...
## $ geometry :sfc_POLYGON of length 8; first list element: List of 1
## ..$ : num [1:3792, 1:2] -77 -77 -77 -77 -77 ...
## ..- attr(*, "class")= chr "XY" "POLYGON" "sfg"
## - attr(*, "sf_column")= chr "geometry"
## - attr(*, "agr")= Factor w/ 3 levels "constant","aggregate",...: NA NA NA NA NA NA NA NA NA ...
## ..- attr(*, "names")= chr "OBJECTID" "WARD" "NAME" "REP_NAME" ...

```

A key attribute of a shapefile is its projection (for more on projections, see lecture notes). To find the projection of a sf file, use the `st_crs()` command. This reports the projection of the file.

```

# make sure it has a projection
st_crs(dc.wards)

```

```
## Coordinate Reference System:
##   EPSG: 4326
##   proj4string: "+proj=longlat +datum=WGS84 +no_defs"
```

What does this information mean? There are (at least) two ways of identifying a projection. The first is the EPSG code. This is a numeric reference for the projection of the file. “EPSG” stands for the European Petroleum Survey Group, which first made this numeric system of reference. They have since been absorbed by the International Association of Oil and Gas Producers (IOGP). The IOGP maintains a registry of projections and associated codes that are very useful to anyone using spatial data.

The second method of identifying a projection is the “proj4string.” This is a set of characters and numbers that uniquely identify a projection. If at all possible, I prefer the EPSG codes, which are much easier to work with.

It’s useful to know the projection of a file if

- you’d like to do some analysis with that file and another shapefile – they should be in the same projection
- to know whether you’ve got the desired projection. For example, some US projections give a flat top to the US, and others give a curved top.

This particular projection is WGS84, a very standard projection maintained by the National Geospatial-Intelligence Agency. It originated in the 1950s, and you can read more [here](#).

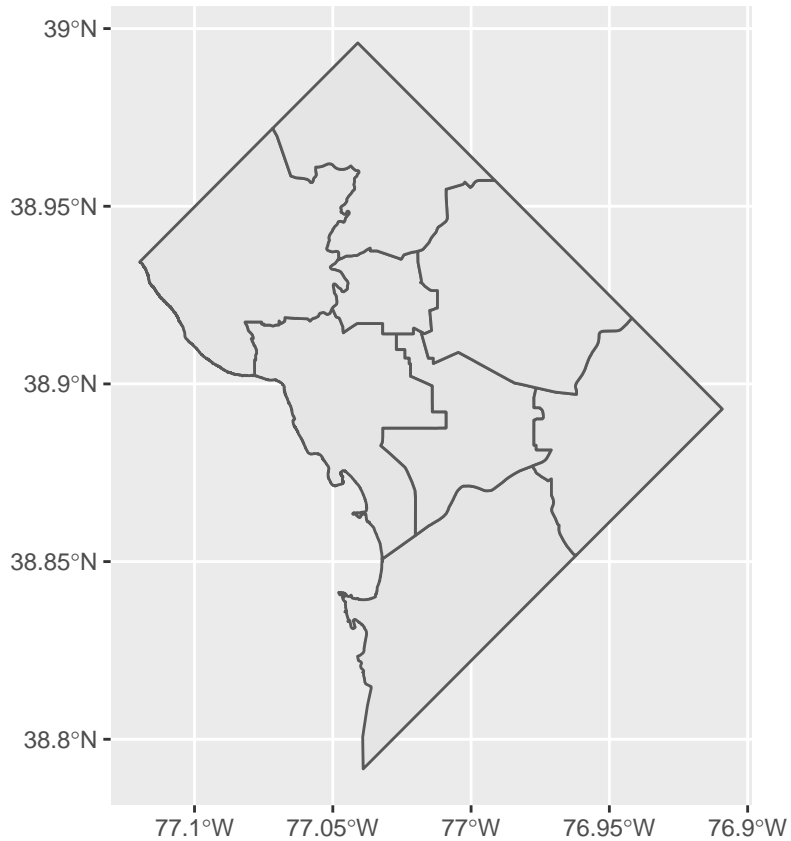
B.3. Plot

There are two major ways to generate a map plot from a `sf` file. You can use the `plot` command or `ggplot`. In the interests of time, we skip `plot` and head straight to `ggplot`, which is much easier to manipulate and customize.

Using `ggplot` here is very similar to what we’ve already done. Instead of `geom_bar()` or `geom_hist()`, we now rely on `geom_sf()`. This command takes a simple feature as the data input and plots the polygons. Because we are plotting only the shape and no variables, there is no `aes()` portion of the command.

The below command is the simplest way to make a map from a “simple feature.” As we’ve done in previous tutorials with graphs, we create an object and then call it to see it. You could simply use `ggplot` without creating an object (though you wouldn’t be able to save it later if you wanted).

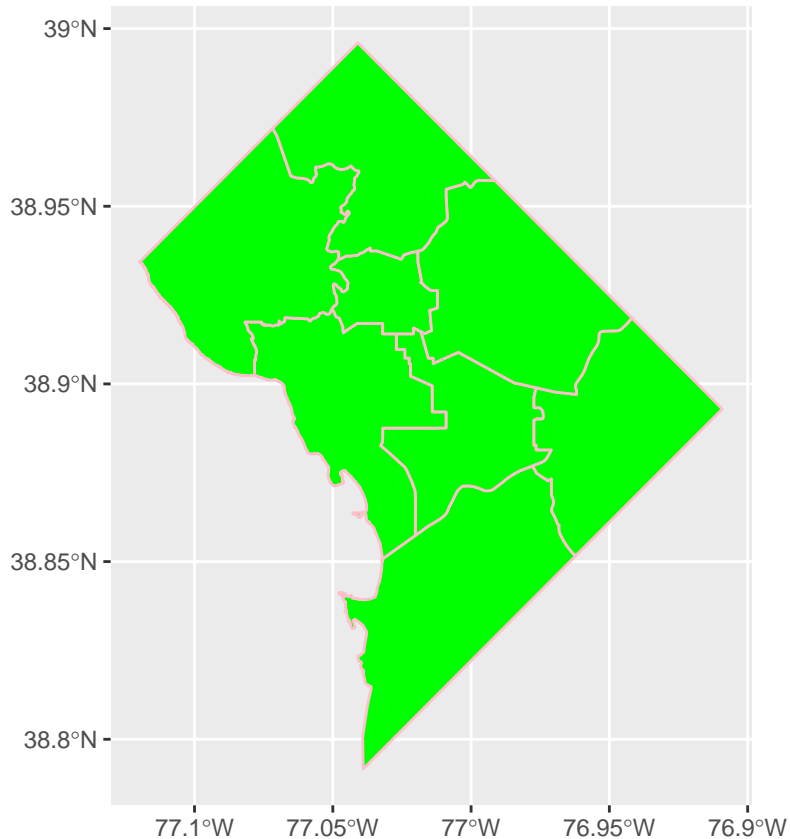
```
# two ways to plot -- second
ward.map <- ggplot() +
  geom_sf(data = dc.wards)
ward.map
```



This is a basic map! You can use this technique to plot any kind of shapefile you can download.

Of course, there are multiple things you might want to change on this map to make it more legible. Some basic modifications include the color of the polygon lines and the polygon fill. Below I set the color of the lines to pink and the fill to green; these commands are both inside the `geom_sf()` command, like graph options are inside `geom_bar()` or `geom_histogram()`.

```
# with color and fill
ward.map <- ggplot() +
  geom_sf(data = dc.wards, color = "pink", fill = "green")
ward.map
```

One key function of a map is to identify things. The map above does not identify wards. Let's fix that. We already have a variable for the ward number, but we have to have a variable that tells R where to put the label. To think about the "where," remember that a map is sort of like a graph, but where the coordinates are latitude and longitude.

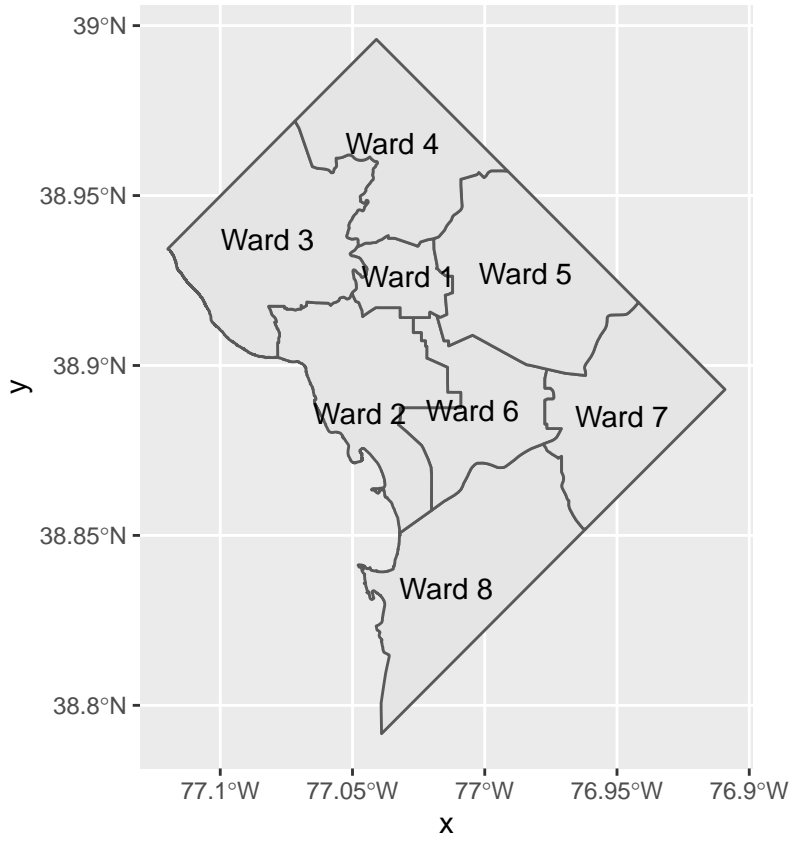
Of course, the current map is polygons, not points. R knows the the locations of the points that bound the polygon, but it doesn't have the latitude and longitude of the center of the polygon, which is where we'd like to put the label.

Helpfully, there is a built-in R command to do this: `geom_sf_text()`. This command puts whatever text you specify in `label = [variable]` at the center of the polygon. (There is also `geom_sf_label()` if you prefer.)¹

```
# plot
ward.map <- ggplot() +
  geom_sf(data = dc.wards) +
  geom_sf_text(data = dc.wards,
              mapping = aes(label=NAME))
ward.map
```

```
## Warning in st_point_on_surface.sfc(sf::st_zm(x)): st_point_on_surface may not
## give correct results for longitude/latitude data
```

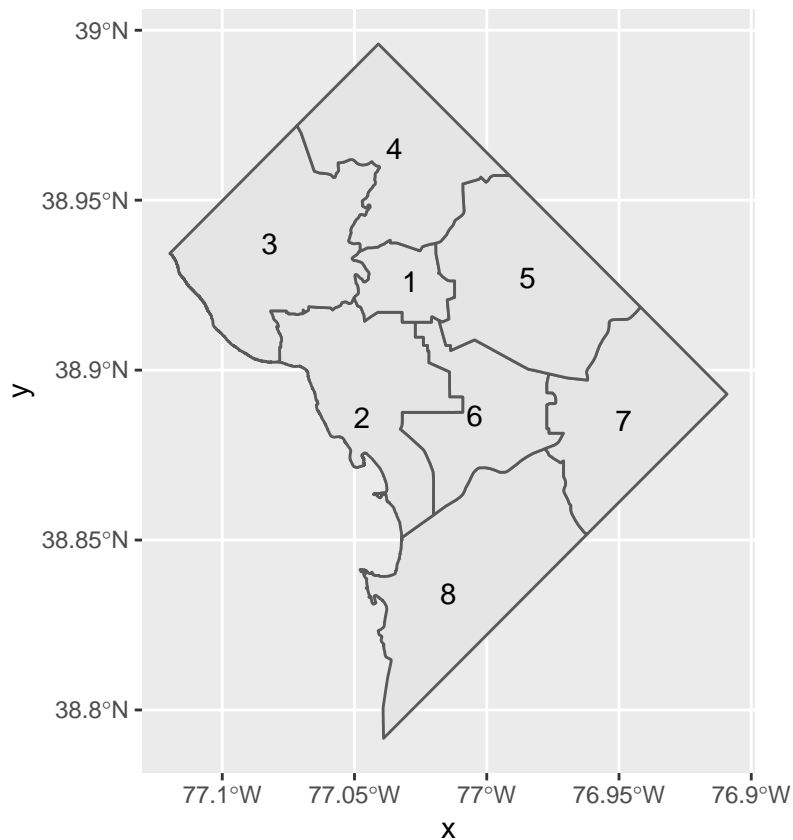
¹In the past, you needed to find the centroid of each polygon and then add this to the dataframe. This may still be useful under certain circumstances. If you are interested, you can find code at <https://yutani.rbind.io/post/geom-sf-text-and-geom-sf-label-are-coming/>.



That looks ok, but you might prefer just the ward number, especially if your title makes it clear that these are all wards. You can change what the label says by changing the variable you associate with the label. Below we label by `WARD`, which is a number for the ward.

```
# or you can do just the number
ward.map <- ggplot() +
  geom_sf(data = dc.wards) +
  geom_sf_text(data = dc.wards,
              mapping = aes(label=WARD))
ward.map
```

```
## Warning in st_point_on_surface.sfc(sf::st_zm(x)): st_point_on_surface may not
## give correct results for longitude/latitude data
```



B.4. Plot two things

One of the most powerful things about the combination of `sf` and `ggplot` is the ability to plot multiple features together. To illustrate this, download the location of DC public school administration [here](#). Again, save as a shapefile somewhere you'll remember. This is a *points* datafile, different from the polygons file we were working with.

Below we load the data, using `st_read()`.

```
# load some points
dcps.pts <- st_read("H:/pppa_data_viz/2019/tutorial_data/lecture05/Public_School_Administration_Points/")

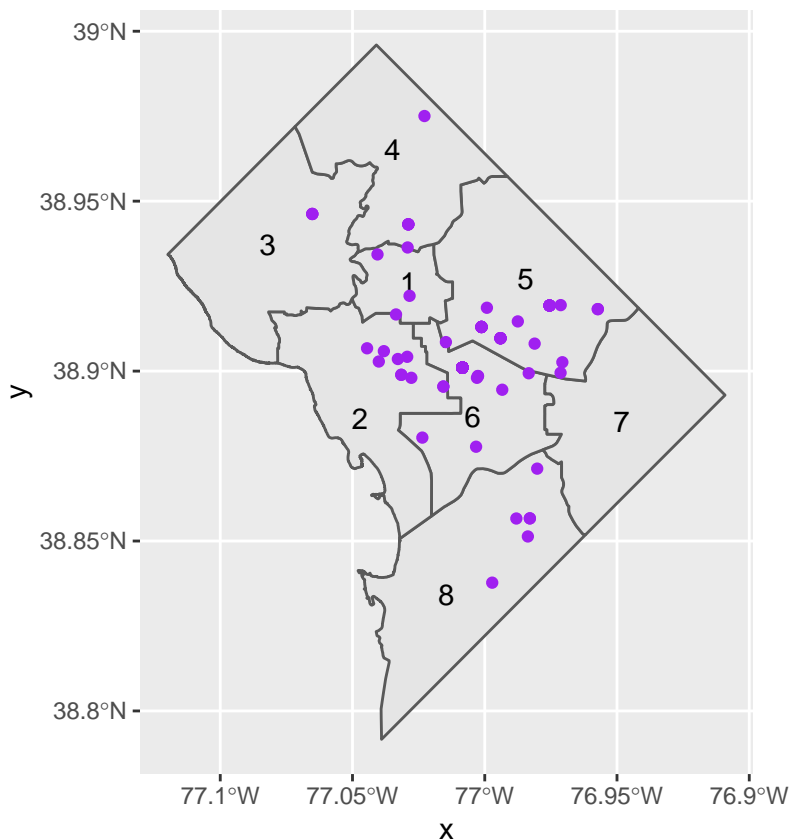
## Reading layer `Public_School_Administration_Points' from data source `H:\pppa_data_viz\2019\tutorial
```

```
## Simple feature collection with 163 features and 8 fields
## geometry type: POINT
## dimension: XY
## bbox: xmin: -77.0652 ymin: 38.83775 xmax: -76.95726 ymax: 38.97505
## epsg (SRID): 4326
## proj4string: +proj=longlat +datum=WGS84 +no_defs
```

With these data loaded, add them to the previous map with an additional `geom_sf()` command, making purple points. Note that you don't need separate plotting commands for different map types. R can figure out which is which.

```
# plot them
ward.map <- ggplot() +
  geom_sf(data = dc.wards) +
  geom_sf_text(data = dc.wards,
              mapping = aes(label=WARD)) +
  geom_sf(data = dcps.pts,
          color = "purple")
ward.map
```

```
## Warning in st_point_on_surface.sfc(sf::st_zm(x)): st_point_on_surface may not
## give correct results for longitude/latitude data
```



Note that to put two maps on the same chart, they need to be in the same projection. R may do this automatically, but you may run into trouble if this fails.

C. Use bigger data

Now we are going to use a larger dataset and try to make a presentation-quality map.

C.1. Load crime data

For this portion of the tutorial, we'll use DC crime data. Download 2018 DC crime data as a shapefile from [here](#); then use `st_read()` to read in the file.

```
# load data
c2018 <- st_read("H:/pppa_data_viz/2019/tutorial_data/lecture05/Crime_Incidents_in_2018/Crime_Incidents_2018.shp")

## Reading layer `Crime_Incidents_in_2018' from data source `H:\pppa_data_viz\2019\tutorial_data\lecture05\Crime_Incidents_in_2018\Crime_Incidents_2018.shp'
## Simple feature collection with 33645 features and 23 fields
## geometry type: POINT
## dimension: XY
## bbox: xmin: -77.11232 ymin: 38.81467 xmax: -76.91002 ymax: 38.9937
## epsg (SRID): 4326
## proj4string: +proj=longlat +datum=WGS84 +no_defs
```

Each row in this dataframe is a crime, with attendant information, including the specific location of the crime.

It is always a good idea to do some quick checks on data quality. The note on DC Open data's page about missing values made me nervous, but using `summary()` to look at the geographic information suggests it is ok.

First we look at the variable definitions:

```
# look at variables
str(c2018)

## Classes 'sf' and 'data.frame': 33645 obs. of 24 variables:
## $ CCN : Factor w/ 33638 levels "07006630","10954295",...: 1 2 3 9 8 12553 12554 12555 12556 12557 ...
## $ REPORT_DAT: Factor w/ 33597 levels "2018-01-01T00:04:32.000Z",...: 17438 7411 18001 492 11323 12888 ...
## $ SHIFT : Factor w/ 3 levels "DAY","EVENING",...: 3 2 3 3 1 3 3 3 3 3 ...
## $ METHOD : Factor w/ 3 levels "GUN","KNIFE",...: 1 3 3 1 3 3 3 1 3 3 ...
## $ OFFENSE : Factor w/ 9 levels "ARSON","ASSAULT W/DANGEROUS WEAPON",...: 4 7 4 4 8 9 5 2 9 8 ...
## $ BLOCK : Factor w/ 7119 levels "0 - 0 BLOCK OF COLUMBUS CIRCLE NE",...: 3779 4432 1880 3114 3177 ...
## $ XBLOCK : num 402412 393499 401872 400393 400863 ...
## $ YBLOCK : num 131645 138307 136822 132326 132294 ...
## $ WARD : Factor w/ 8 levels "1","2","3","4",...: 8 2 6 8 8 8 5 2 7 6 ...
## $ ANC : Factor w/ 40 levels "1A","1B","1C",...: 37 9 26 38 36 38 23 6 33 30 ...
## $ DISTRICT : Factor w/ 7 levels "1","2","3","4",...: 7 2 5 7 7 7 5 3 5 1 ...
## $ PSA : Factor w/ 57 levels "101","102","103",...: 51 14 41 52 52 56 37 18 41 1 ...
## $ NEIGHBORHO : Factor w/ 39 levels "Cluster 1","Cluster 10",...: 30 34 16 31 31 33 17 36 18 38 ...
## $ BLOCK_GROU : Factor w/ 449 levels "000100 1","000100 2",...: 261 5 311 252 257 430 447 210 315 190 ...
## $ CENSUS_TRA : Factor w/ 179 levels "000100","000201",...: 99 2 118 94 97 172 179 75 119 66 ...
## $ VOTING_PRE : Factor w/ 143 levels "Precinct 1","Precinct 10",...: 19 100 124 23 22 28 113 46 123 1 ...
## $ LATITUDE : num 38.9 38.9 38.9 38.9 38.9 ...
## $ LONGITUDE : num -77 -77.1 -77 -77 -77 ...
## $ BID : Factor w/ 10 levels "ADAMS MORGAN",...: NA NA NA 2 NA NA NA NA NA 8 ...
## $ START_DATE: Factor w/ 33517 levels "1985-01-11T03:56:21.000Z",...: 5 7166 8 29 11459 13047 13054 1 ...
## $ END_DATE : Factor w/ 27256 levels "1985-01-11T04:00:39.000Z",...: 4 6688 7 NA NA 10770 NA 10773 NA ...
## $ OBJECTID : num 2.59e+08 2.59e+08 2.59e+08 2.59e+08 2.59e+08 ...
## $ OCTO_RECOR: Factor w/ 33645 levels "07006630-01",...: 1 2 3 9 8 12559 12560 12561 12562 12563 ...
## $ geometry :sfc_POINT of length 33645; first list element: 'XY' num -77 38.9
```

```
## - attr(*, "sf_column")= chr "geometry"
## - attr(*, "agr")= Factor w/ 3 levels "constant","aggregate",...: NA NA NA NA NA NA NA NA NA NA ...
## ..- attr(*, "names")= chr "CCN" "REPORT_DAT" "SHIFT" "METHOD" ...
```

Then I check out a few specific variables.

```
# size and missings
dim(c2018)
```

```
## [1] 33645 24
```

```
summary(c2018$LATITUDE)
```

```
## Min. 1st Qu. Median Mean 3rd Qu. Max.
## 38.81 38.89 38.91 38.91 38.92 38.99
```

```
summary(c2018$LONGITUDE)
```

```
## Min. 1st Qu. Median Mean 3rd Qu. Max.
## -77.11 -77.03 -77.01 -77.01 -76.99 -76.91
```

```
table(c2018$OFFENSE)
```

```
##
## ARSON ASSAULT W/DANGEROUS WEAPON
## 5 1668
## BURGLARY HOMICIDE
## 1415 160
## MOTOR VEHICLE THEFT ROBBERY
## 2388 2019
## SEX ABUSE THEFT F/AUTO
## 275 11574
## THEFT/OTHER
## 14141
```

```
# nothing seems to have NAs
```

There does not seem to be an overwhelming number of missing values (NAs).

C.2. Making legible maps from these data

There are so many crimes in these data that a picture that maps them all is not a good idea. So we make a marker for the violent crimes (arson, assault, homicide, robbery and sex abuse) only. We use a `ifelse()` command to discriminate between the two types. After creating the new variable `ctype`, I use the `table()` command to check whether I've put the right crimes in each group.

Note that I use some shorthand code so I don't have to write six equalities. Writing `c2018$OFFENSE %in% c("ARSON","ASSAULT W/DANGEROUS WEAPON")` is equivalent to writing `c2018$OFFENSE == "ARSON" | c2018$OFFENSE == "ASSAULT W/DANGEROUS WEAPON"`. The first is shorter and easier to debug.

```
# wrong!
c2018$ctype <- ifelse(test = c2018$OFFENSE %in% c("ARSON","ASSAULT W/DANGEROUS WEAPON",
"HOMICIDE", "ROBBERY", "SEX ABUSE"),
yes = 1,
no = 0)
```

It is always a good idea to check your work. As we see below, using a check with the `table()` command, my coding did not catching all violent crimes.

```
table(c2018$ctype, c2018$OFFENSE)
```

```
##
##      ARSON ASSAULT W/DANGEROUS WEAPON BURGLARY HOMICIDE MOTOR VEHICLE THEFT
##    0      0                                1668    1415         0           2388
##    1      5                                0         0         160           0
##
##      ROBBERY SEX ABUSE THEFT F/AUTO THEFT/OTHER
##    0          0          0          11574         14141
##    1      2019          275           0           0
```

The “assault with a dangerous weapon” should have been a violent crime, but it is not. An extra space above that caused trouble. I fix and try again:

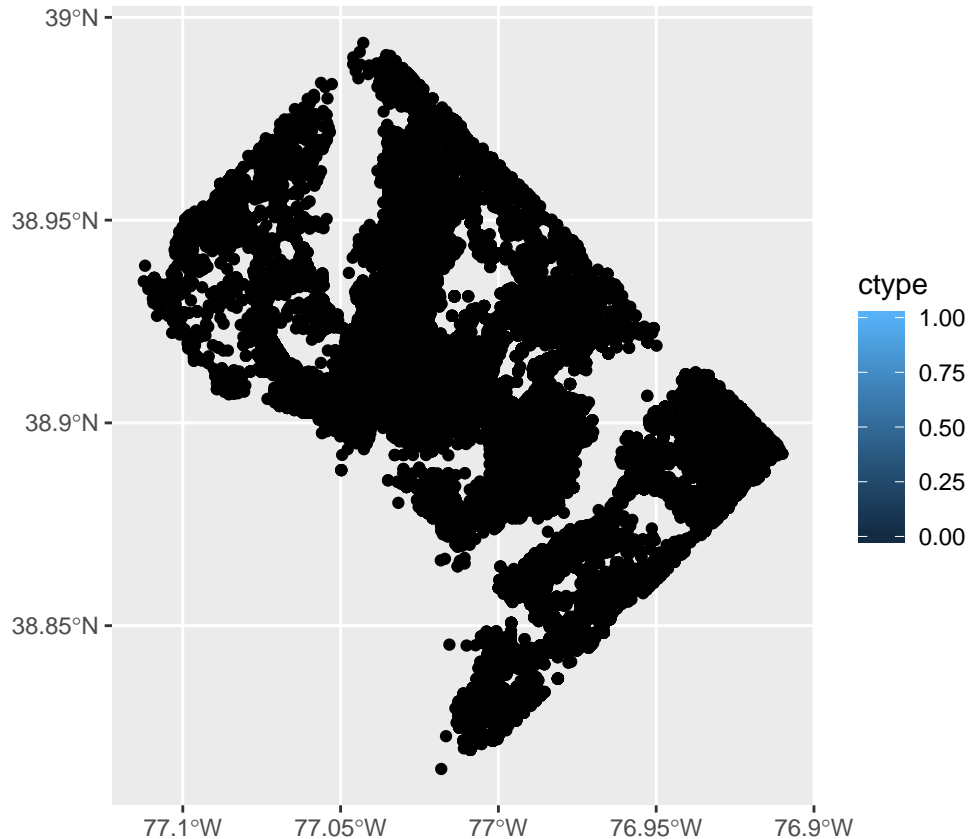
```
# better!
c2018$ctype <- ifelse(test = c2018$OFFENSE %in% c("ARSON", "ASSAULT W/DANGEROUS WEAPON",
                                                "HOMICIDE", "ROBBERY", "SEX ABUSE"),
                      yes = 1,
                      no = 0)
table(c2018$ctype, c2018$OFFENSE)
```

```
##
##      ARSON ASSAULT W/DANGEROUS WEAPON BURGLARY HOMICIDE MOTOR VEHICLE THEFT
##    0      0                                0    1415         0           2388
##    1      5                                1668     0         160           0
##
##      ROBBERY SEX ABUSE THEFT F/AUTO THEFT/OTHER
##    0          0          0          11574         14141
##    1      2019          275           0           0
```

Now this looks ok. All crimes that I listed as violent show up as having `c2018$ctype == 1`.

Another way to check the data is to plot them. Do this below. But if your computer chokes, just ignore this step:

```
# plot
crimeo <- ggplot() +
  geom_sf(data = c2018, aes(fill = ctype))
crimeo
```



This is so many points that it's hard to make sense of the map.

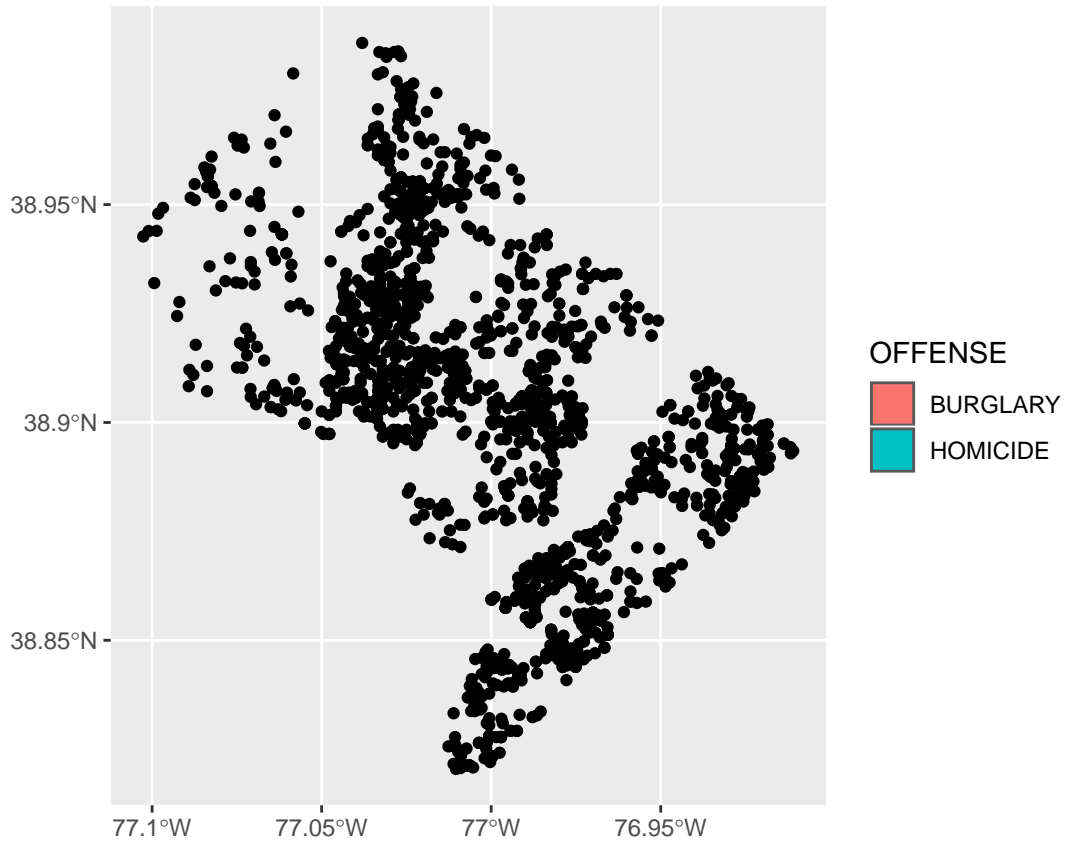
One way to have a map that is legible is to focus on crimes with fewer occurrences. To this end, we subset to a smaller dataframe with just homicides and burglaries. We use the type of subsetting we've already learned. Recall that the statement `(c2018$OFFENSE == "HOMICIDE" | c2018$OFFENSE == "BURGLARY")` is true if the offense is *either* a homicide or a burglary. The `|` operator means "or" (in R and many other languages).

```
# make smaller dataframe
vc2018 <- c2018[which(c2018$OFFENSE == "HOMICIDE" | c2018$OFFENSE == "BURGLARY"),]
```

The `vc2018` dataframe should have fewer observations than `c2018` – you can check that it does with `dim()`.

Now let's plot homicides and burglaries. The code is similar to before, but relies on different data. Notice the change in the `data =` input. Analogous to what we do with bar graphs, we use `fill` for the offense. Notice that this is inside the `aes()` command.

```
# lets just do homicides and burglary -- bad
vc <- ggplot() +
  geom_sf(data = vc2018,
          mapping = aes(fill = OFFENSE))
vc
```

This was not a good idea! From the legend, R knows that we're trying to use these two types, but there's nothing to see in the plot. Rather than `fill=`, use `color=`.

```
# lets just do homicides and burglary -- better
vc <- ggplot() +
  geom_sf(data = vc2018,
          mapping = aes(color = OFFENSE))
vc
```



But the legend now looks a little wacky. Use both `color` and `fill` to get something reasonable looking.

```
# lets just do homicides and burglary -- best
vc <- ggplot() +
  geom_sf(data = vc2018,
          mapping = aes(color = OFFENSE, fill = OFFENSE))
vc
```



Of course, this map still looks ugly. I very much like the map [here](#), made entirely in `ggplot` and `sf`.

To clean up this map, I've mostly copied Timo's work, particularly the theme elements of the map. See the next command below.

I'd like to also add ward boundaries to the map. Here is a first attempt.

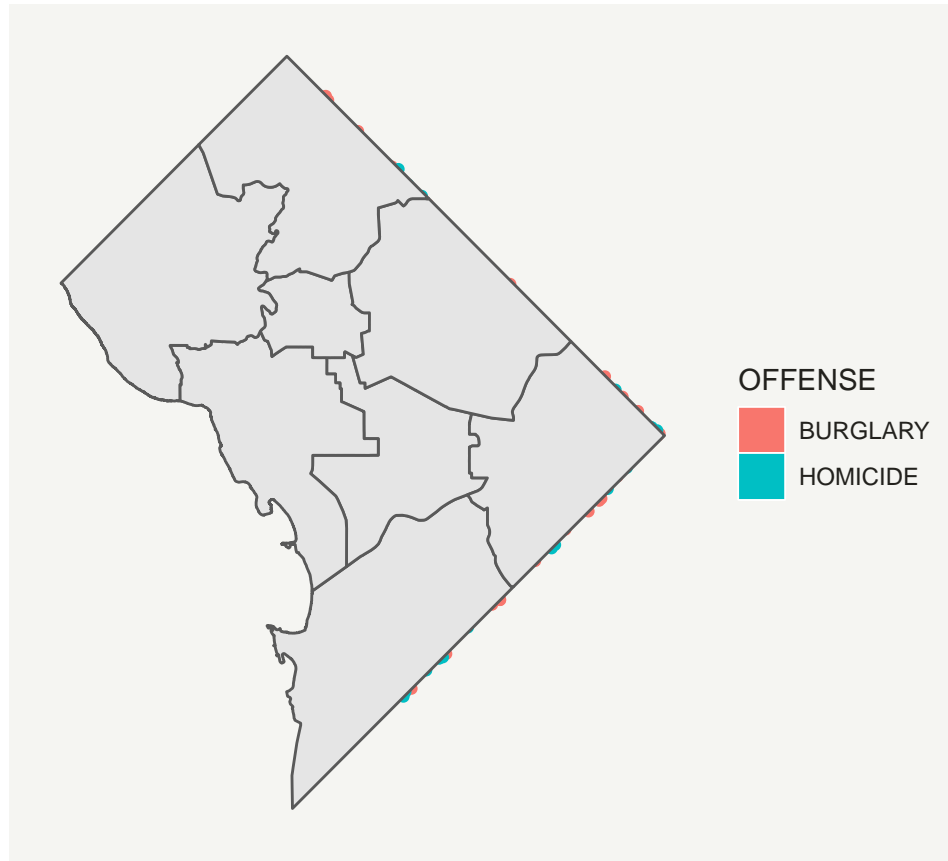
```
# make it look decent -- but wrong layer on top!
vc <- ggplot() +
  geom_sf(data = vc2018,
          mapping = aes(color = OFFENSE, fill = OFFENSE)) +
  geom_sf(data = dc.wards) +
  theme(
    text = element_text(color = "#22211d"),
    axis.line = element_blank(),
    axis.text.x = element_blank(),
    axis.text.y = element_blank(),
    axis.ticks = element_blank(),
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
```

```

panel.grid.minor = element_blank(),
plot.background = element_rect(fill = "#f5f5f2", color = NA),
panel.background = element_rect(fill = "#f5f5f2", color = NA),
legend.background = element_rect(fill = "#f5f5f2", color = NA)
)

```

vc



This looks bad! The code above does not give a good result because the wards are on top of the points. Note that R draws the maps in the order you call them, so that the last one is on top.

We can fix this ordering issue by switching the order of the layering. By putting the wards first in the command, they go first on the map. Then points (vc2018) on top.

```

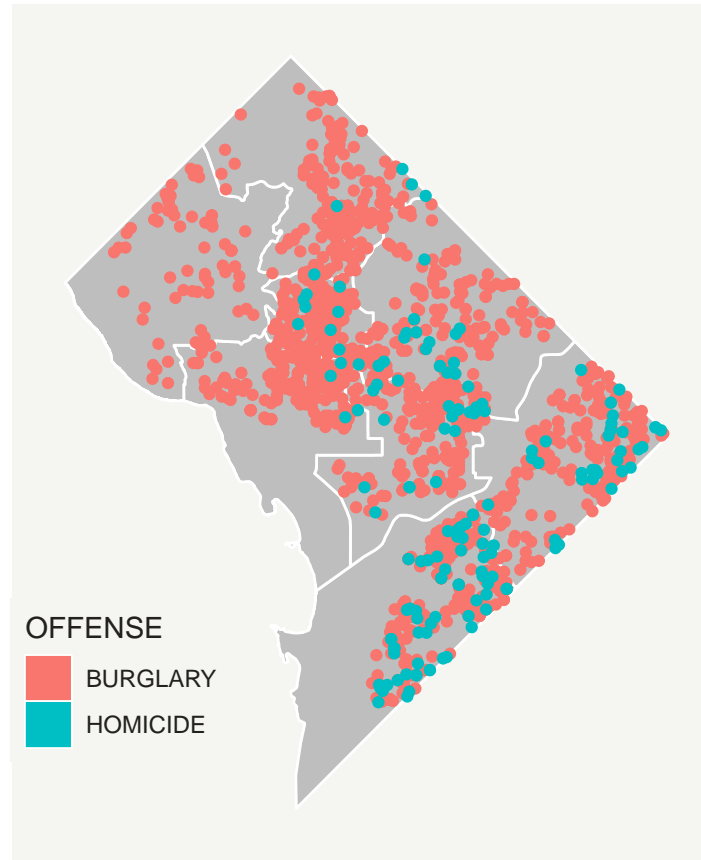
# make it look decent -- fixed layer issue
# substantially copied from
# https://timogrossenbacher.ch/2016/12/beautiful-thematic-maps-with-ggplot2-only/
vc <- ggplot() +
  geom_sf(data = dc.wards, color = "white", fill = "grey") +
  geom_sf(data = vc2018, aes(color = OFFENSE, fill = OFFENSE)) +
  theme(
    text = element_text(color = "#22211d"),
    axis.line = element_blank(),
    axis.text.x = element_blank(),
    axis.text.y = element_blank(),
    axis.ticks = element_blank(),
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
  )

```

```

plot.background = element_rect(fill = "#f5f5f2", color = NA),
panel.background = element_rect(fill = "#f5f5f2", color = NA),
panel.grid = element_line(color = "#f5f5f2"),
legend.background = element_rect(fill = "#f5f5f2", color = NA),
legend.position = c(0.13,0.2)
)
vc

```



Now save this file using `ggsave()` as we learned in Tutorial 4.

Before doing that, I make a variable (or a 1x1 matrix, if you prefer) that has today's date in it, in numeric form (e.g., 20190225). I like to always put the date in the name of anything I save so that when I work on it again I will not save over the older version.

R has a built in function to deliver today's date (`Sys.Date()`). I use the substring command (see lecture 4 notes) to extract the parts we need. I then use `paste0` to put the parts together. This command puts together all the text bits you put into it, with no spaces in between (the 0 part). You can look at all the parts separately if you'd like to better understand what's going on.

```

# save using ggsave
# todays date
Sys.Date()

```

```
## [1] "2020-02-19"
```

```

dateo <- paste0(substr(Sys.Date(),1,4),
                substr(Sys.Date(),6,7),
                substr(Sys.Date(),9,10))

```

```
dateo
```

```
## [1] "20200219"
```

Now with this date in hand, I make the filename with the date in it. If you don't specify the full file path, R saves your file in the "current working directory." The current working directory may *not* be the directory your program is in. (You can see what directory you're in with the `getwd()` command).

The `ggsave()` code has one modification from Tutorial 4. Before the `ggsave()` command, I create a variable that has the filename for the file I'd like to save. Sometimes this makes the programming clearer. So, instead of the long filename in the `filename =` portion of the `ggsave`, I put the variable (`patho`) that I just created.

```
fn <- paste0(dateo, "_violent_crime_2018.jpg")
patho <- paste0("H:/pppa_data_viz/2020/tutorials/tutorial_05/", fn)
ggsave(filename = patho,
        plot = vc,
        device = "jpg",
        dpi = 300)
```

```
## Saving 6.5 x 4.5 in image
```

D. Putting map data together with other data

Part of the value data in a spatial format is your ability to combine data spatially. This means you can combine data not just by merging variables, but by saying "what polygon does this point fall in?" or "which polygons does this polygon touch?"

We now perform such a spatial analysis. For our example, we'll use the crime data and find the block group in which each crime takes place.

This could be useful for a number of reasons. For example, suppose you want to know if poor neighborhoods have more crimes, or if neighborhoods with more shade trees have more crimes. For either of these questions, you need to know the attributes of neighborhoods in which crimes take place. To do this, you need to know a neighborhood ID so you can merge in attributes.

In this tutorial, we use neighborhood information to compute crime rates: crimes divided by population. We already have crimes, but we have to associate each crime with an area, and we need to know the population of each of these areas.

To give an overview, we need to

- find the block group in which each crime is located
- find the total number of crimes in each block group by type → block group level data
- add the block group population
- calculate a rate

D.1. Find the block group of each point

We begin by loading the block group [map](#) I downloaded from the DC open data website; be sure to download the shapefile.

Load the block group data – note that it is a shapefile so we use `st_read()` and see what variables it has.

```
# load block group map
bg2010 <- st_read("H:/pppa_data_viz/2019/tutorial_data/lecture05/Census_Block_Groups__2010/Census_Block
```

```
## Reading layer `Census_Block_Groups_2010' from data source `H:\pppa_data_viz\2019\tutorial_data\lect
## Simple feature collection with 450 features and 54 fields
## geometry type: POLYGON
## dimension: XY
## bbox: xmin: -77.11976 ymin: 38.79165 xmax: -76.9094 ymax: 38.99581
## epsg (SRID): 4326
## proj4string: +proj=longlat +datum=WGS84 +no_defs
```

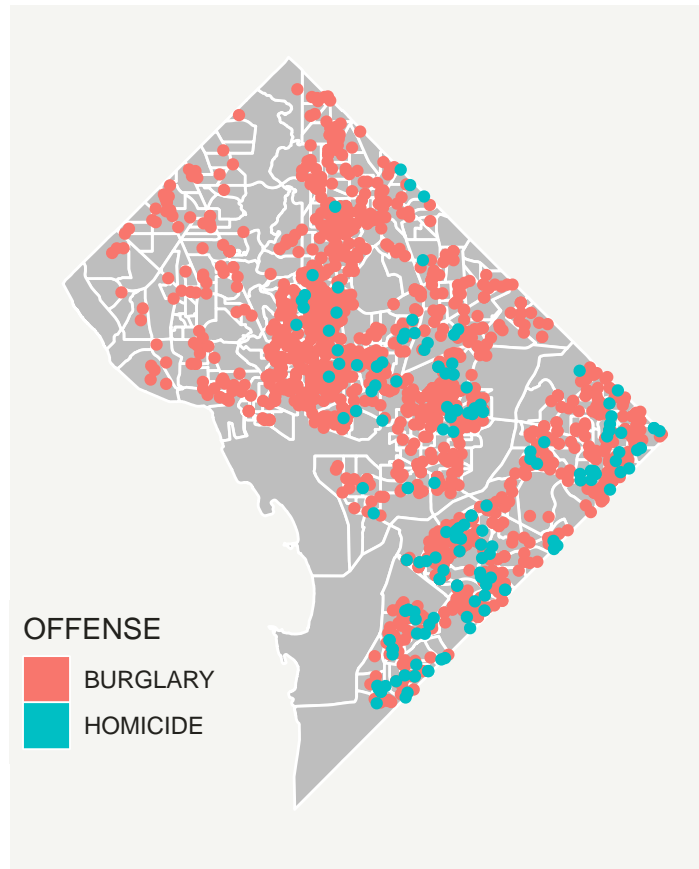
```
names(bg2010)
```

```
## [1] "OBJECTID" "TRACT" "BLKGRP" "GEOID" "P0010001"
## [6] "P0010002" "P0010003" "P0010004" "P0010005" "P0010006"
## [11] "P0010007" "P0010008" "OP000001" "OP000002" "OP000003"
## [16] "OP000004" "P0020002" "P0020005" "P0020006" "P0020007"
## [21] "P0020008" "P0020009" "P0020010" "OP000005" "OP000006"
## [26] "OP000007" "OP000008" "P0030001" "P0030003" "P0030004"
## [31] "P0030005" "P0030006" "P0030007" "P0030008" "OP000009"
## [36] "OP000010" "OP000011" "OP000012" "P0040002" "P0040005"
## [41] "P0040006" "P0040007" "P0040008" "P0040009" "P0040010"
## [46] "OP000013" "OP000014" "OP000015" "OP000016" "H0010001"
## [51] "H0010002" "H0010003" "SHAPE_Leng" "SHAPE_Area" "geometry"
```

Let's plot it with the crimes to make sure that the maps are what we think they are.

```
# look at it with homicides and burglaries
```

```
cbg <- ggplot() +
  geom_sf(data = bg2010, color = "white", fill = "grey") +
  geom_sf(data = vc2018, aes(color = OFFENSE, fill = OFFENSE)) +
  theme(
    text = element_text(color = "#22211d"),
    axis.line = element_blank(),
    axis.text.x = element_blank(),
    axis.text.y = element_blank(),
    axis.ticks = element_blank(),
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    plot.background = element_rect(fill = "#f5f5f2", color = NA),
    panel.background = element_rect(fill = "#f5f5f2", color = NA),
    panel.grid = element_blank(),
    legend.background = element_rect(fill = "#f5f5f2", color = NA),
    legend.position = c(0.13,0.2)
  )
cbg
```



This quick look suggests that all crimes are inside a block group. Now we need to find which block group. To do this, we use a command called `st_intersection()`. When the shapefiles are [points, polygons] like the ones in the call below, R returns a points dataframe with information on the block group into which each point falls.

If this seems confusing, please go back to the lecture notes and look at the example about this type of intersection.

Let's also make the intersection easier on ourselves by making a smaller version of the block group data (`bg2010.small`) that has just the block group id variables – that's all we want to add to the points and it is helpful to reduce the number of variables we need to keep track of.

```
# find which crimes are in which block groups
bg2010.small <- bg2010[,c("TRACT", "BLKGRP")]
dim(bg2010.small)
```

```
## [1] 450 3
```

Finally, before intersecting, make sure both simple features have the same projection. If the files do not, the intersection will either fail or give you garbage.

```
st_crs(vc2018)
```

```
## Coordinate Reference System:
## EPSG: 4326
## proj4string: "+proj=longlat +datum=WGS84 +no_defs"
```

```
st_crs(bg2010.small)
```

```
## Coordinate Reference System:
```



```
## EPSG: 4326
## proj4string: "+proj=longlat +datum=WGS84 +no_defs"
```

This looks good. But if you want to be clever, you can even test their equality:

```
proj.test <- "nada"
proj.test <- ifelse(test = (st_crs(vc2018) = st_crs(bg2010.small)),
  yes = "yes",
  no = "no")
print(paste0("Are the projections the same? ", proj.test))
```

```
## [1] "Are the projections the same? yes" "Are the projections the same? NA"
```

Now that we've confirmed the two projections are the same, use this smaller dataframe to do the points and polygons intersection:

```
# find which crimes are in which block groups
cbg <- st_intersection(vc2018,bg2010.small)
```

```
## although coordinates are longitude/latitude, st_intersection assumes that they are planar
## Warning: attribute variables are assumed to be spatially constant throughout all
## geometries
```

It is critically important to check the output.

```
head(cbg)
```

```
## Simple feature collection with 6 features and 26 fields
## geometry type: POINT
## dimension: XY
## bbox: xmin: -77.0734 ymin: 38.90259 xmax: -77.05759 ymax: 38.91258
## epsg (SRID): 4326
## proj4string: +proj=longlat +datum=WGS84 +no_defs
## CCN REPORT_DAT SHIFT METHOD OFFENSE
## 1826 18064713 2018-04-23T13:04:28.000Z DAY OTHERS BURGLARY
## 11452 18134729 2018-08-14T03:10:42.000Z MIDNIGHT OTHERS BURGLARY
## 26674 18042061 2018-03-15T16:09:21.000Z EVENING OTHERS BURGLARY
## 30745 18026424 2018-02-16T14:31:32.000Z DAY OTHERS BURGLARY
## 30750 18026435 2018-02-16T15:42:03.000Z EVENING OTHERS BURGLARY
## 23472 18136654 2018-08-17T04:38:28.000Z MIDNIGHT OTHERS BURGLARY
## BLOCK_XBLOCK_YBLOCK_WARD_ANC
## 1826 3100 - 3199 BLOCK OF K STREET NW 394626 137194 2 2E
## 11452 3036 - 3099 BLOCK OF M STREET NW 394737 137483 2 2E
## 26674 3000 - 3099 BLOCK OF N STREET NW 394778 137665 2 2E
## 30745 2800 - 2899 BLOCK OF PENNSYLVANIA AVENUE NW 395005 137471 2 2E
## 30750 2800 - 2899 BLOCK OF PENNSYLVANIA AVENUE NW 395005 137471 2 2E
## 23472 3700 - 3799 BLOCK OF RESERVOIR ROAD NW 393634 138304 2 2E
## DISTRICT_PSA_NEIGHBORHO_BLOCK_GROU_CENSUS_TRA_VOTING_PRE_LATITUDE
## 1826 2 206 Cluster 4 000100 4 000100 Precinct 5 38.90258
## 11452 2 206 Cluster 4 000100 4 000100 Precinct 5 38.90519
## 26674 2 206 Cluster 4 000100 4 000100 Precinct 5 38.90683
## 30745 2 206 Cluster 4 000100 4 000100 Precinct 5 38.90508
## 30750 2 206 Cluster 4 000100 4 000100 Precinct 5 38.90508
## 23472 2 206 Cluster 4 000201 1 000201 Precinct 6 38.91258
## LONGITUDE BID START_DATE END_DATE
## 1826 -77.06195 GEORGETOWN 2018-04-23T10:56:35.000Z <NA>
## 11452 -77.06068 GEORGETOWN 2018-08-13T22:52:11.000Z 2018-08-14T00:19:42.000Z
```

```
## 26674 -77.06021 <NA> 2018-03-14T03:48:45.000Z 2018-03-14T03:49:12.000Z
## 30745 -77.05759 GEORGETOWN 2018-02-16T01:25:28.000Z 2018-02-16T02:30:15.000Z
## 30750 -77.05759 GEORGETOWN 2018-02-15T17:30:53.000Z 2018-02-16T08:15:53.000Z
## 23472 -77.07340 <NA> 2018-08-17T03:36:29.000Z 2018-08-17T03:46:04.000Z
## OBJECTID OCTO_RECOR ctype TRACT BLKGRP geometry
## 1826 259183436 18064713-01 0 000100 4 POINT (-77.06196 38.90259)
## 11452 259193062 18134729-01 0 000100 4 POINT (-77.06068 38.9052)
## 26674 259240847 18042061-01 0 000100 4 POINT (-77.06021 38.90684)
## 30745 259245997 18026424-01 0 000100 4 POINT (-77.05759 38.90509)
## 30750 259246002 18026435-01 0 000100 4 POINT (-77.05759 38.90509)
## 23472 259224604 18136654-01 0 000201 1 POINT (-77.0734 38.91258)
```

This file should have the same number of rows as the crime points datafile. It should have variables from both the crime points file and the block group file. Does this seem right?

Now I can start to find the average number of crimes per person. This is a multi-step process:

- find the block group in which each crime is located (done)
- find the total number of crimes in each block group for each type → block group level data
- add the block group population
- calculate a rate

D.2. Total number of crimes per block group

The next step in this process of finding a crime rate by block group is to find the total number of crimes by type in each block group. Remember that the current dataset is at the crime level – one observation per crime. We want a dataset at the block group/offense level – so we know how many offenses of each type are committed in each block group in 2018.

We want R to just count the number of observations by block group and offense. We do this by grouping the data by tract/block group/offense, and then summarizing (counting) the number of observations in each group (`n()`).

```
# first count number of crimes by type by block group
require(dplyr)
dim(cbg)
```

```
## [1] 1567 27
```

```
cbg <- group_by(.data = cbg, TRACT, BLKGRP, OFFENSE)
cbgs <- summarize(.data = cbg, incidents = n())
```

This new dataframe should have two rows for each block group, if each block group has both types of crimes (homicides and burglaries). Since not all block groups have homicides (thank goodness), the total number of observations should be less than the total number of block groups times two. Under NO circumstances should it ever be more. So a good first initial check on this dataframe (`cbgs`) is to see how many rows it has relative to the block group file. It should have more, but never more than twice as much.

```
# total number of observations and variables in new block group/crime file
dim(cbgs)
```

```
## [1] 478 5
```

```
# total number of observations and variables in block group file
dim(bg2010.small)
```

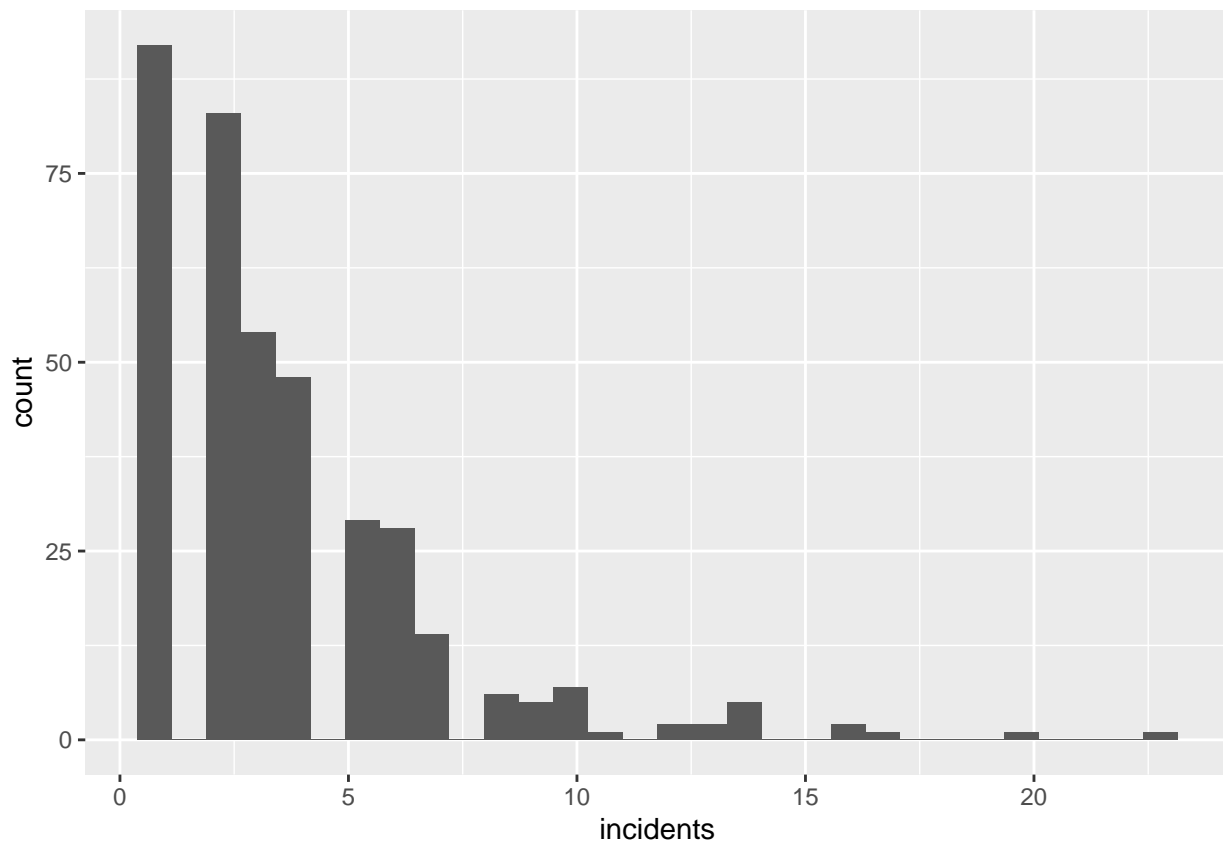
```
## [1] 450 3
```

Things look ok.

Let's check this work a different way by making a quick histogram of burglaries (homicides, thankfully, are more rare, so they are less amenable to this kind of check). Note that this shows only block groups with at least one burglary. (We could fix this with what we do later in this step.)

```
# make a histogram of the distribution of burglaries
burg.hist <- ggplot() +
  geom_histogram(data = cbgs[which(cbgs$OFFENSE == "BURGLARY"),],
                mapping = aes(x = incidents))
burg.hist

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Look at these numbers. Do these seem like plausible numbers for burglaries in a given year in a block group?

So now you have successfully created a block group/offense level dataframe (one row per block group and offense type).

D.3. Add block group population

Now we are ready to undertake the third step, which is adding block group population.

Remember, the steps are * find the block group in which each crime is located (done) * find the total number of crimes in each block group for each type → block group level data (done) * add the block group population * calculate a rate

We do this because it's useful to know not just the level of crime, but the crime rate. We find a crime rate by

dividing the number of crimes by the resident population (what is the right denominator is a big issue for criminologists; for now we'll suffice with resident population).

To make a rate, we need the resident population of each block group. I know that the block group shapefile has a population variable. It's poorly labeled online, but I know from other work that P0010001 is actually total population.

We cannot merge a spatial dataframe (`cbgs`) and another spatial dataframe (`bgs.2010`) R would be confused and not know which shapes to use.

To solve this problem, we'll make the block group file a regular dataframe, rather than a simple feature. We do this by setting the geometry part of the file to "NULL." We also limit the variables in the block group file for ease of use.

```
# need to merge in block group population
# i know that total population is P0010001
bg2010.pop <- bg2010[,c("TRACT", "BLKGRP", "P0010001")]
st_geometry(bg2010.pop) <- NULL
```

Now we have a new dataframe called `bg2010.pop` that has no spatial information and three variables. Next we merge with dataframe with the `cbg` dataframe by two variables: tract and block group. (If you were using more than just one state's worth of data, you would need to include the state ID as well; tract numbers repeat across states.) We specify the `by =` to have two variables using this `c()` notation. We use `all = TRUE` to keep all observations from both dataframes. This has the effect of keeping all block groups, meaning that we have block groups even when they have no homicides or burglaries. This allows us to calculate crime rates of zero.

After this merge (as with any merge), we check the size of the output dataframe relative to the size of the input dataframes.

```
cbgs2 <- merge(x = cbgs,
              y = bg2010.pop,
              by = c("TRACT", "BLKGRP"),
              all.x = TRUE)
dim(cbgs)
```

```
## [1] 478  5
```

```
dim(bg2010.pop)
```

```
## [1] 450  3
```

```
dim(cbgs2)
```

```
## [1] 478  6
```

```
# check result of merge
summary(cbgs2$P0010001)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##         33     930    1292    1382    1741    3916
```

In the homework, you'll report how many block groups have no reported crime in 2018.

D.4. Find crime rate

Now we are finally ready to make a crime rate. We have

- found the block group in which each crime is located
- found the total number of crimes in each block group for each type → block group level data

- added the block group population .. and we need only
- calculate a rate

At the moment, if a block group has no homicide or burglaries, it has a missing value for the variable `incidents`. If we used this in a calculation, we'd get a missing value for the crime rate (`NA/population = NA`).

```
# set incidents equal to zero if missing
summary(cbgs2$incidents)

##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  1.000  1.000   2.000   3.278  4.000  23.000

cbgs2$incidents <- ifelse(test = is.na(cbgs2$incidents) == TRUE,
                          yes = 0,
                          no = cbgs2$incidents)
```

Now that we've fixed the NA issue, let's calculate a crime rate. Because the rate per person is quite small (which is good), we calculate a rate per 100 people to make the number legible.

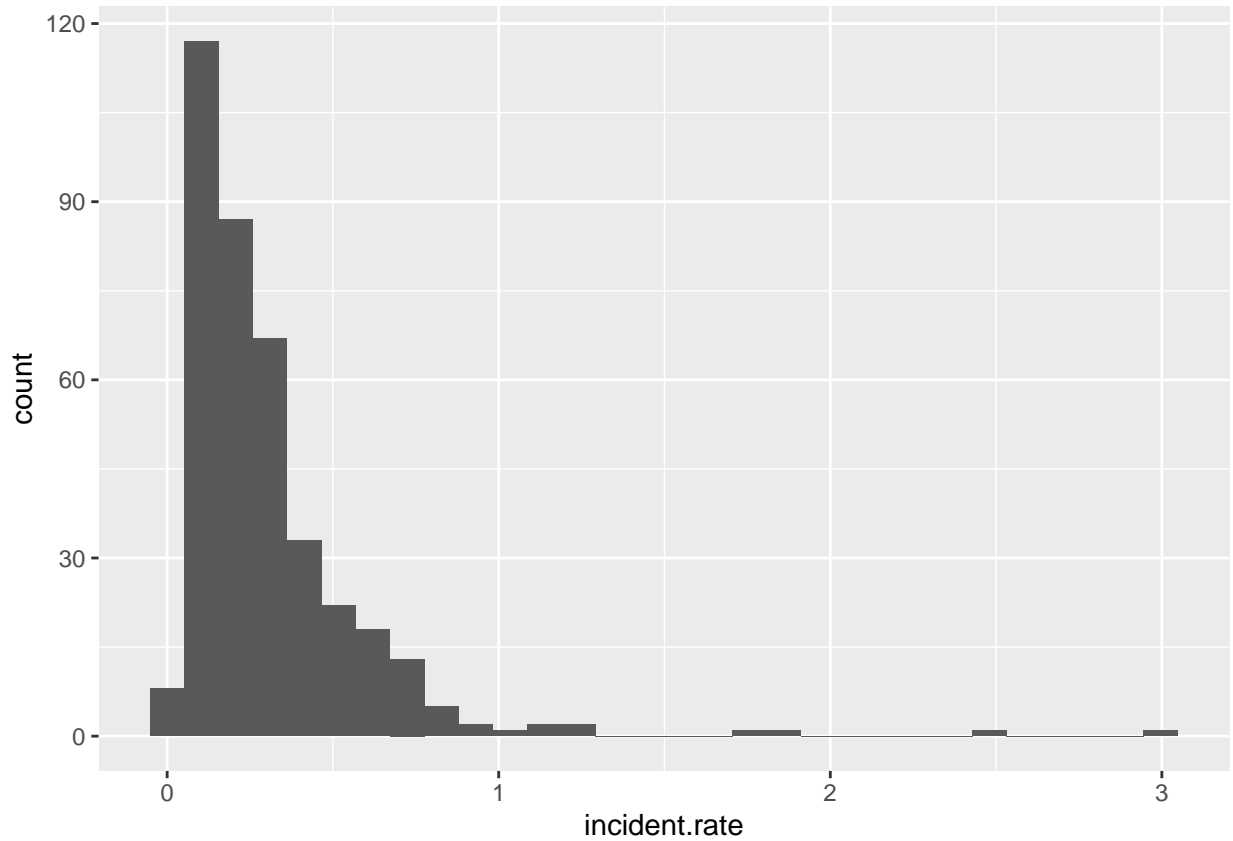
```
# create crime rate
cbgs2$incident.rate = (cbgs2$incidents / (cbgs2$P0010001/100))
summary(cbgs2$incident.rate)
```

```
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.03336 0.10564 0.18859 0.27505 0.33264 3.03030
```

To get a sense of this distribution, make a histogram with these rates.

```
# make the histogram
rburg.hist <- ggplot() +
  geom_histogram(data = cbgs2[which(cbgs2$OFFENSE == "BURGLARY"),],
                mapping = aes(x = incident.rate))
rburg.hist
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



E. Homework

1. How many block groups have no homicides, and how many have no burglaries in 2018? What share of all block groups are these two numbers?
2. Make a bar chart that shows the population density (people / area) by ward. You may wish to use `st_area()` to find the area of each ward. Population is already in the ward dataframe.
3. Get two other maps not used in this tutorial. Layer them and make them look decent.