Assignment4

Zach Rippas

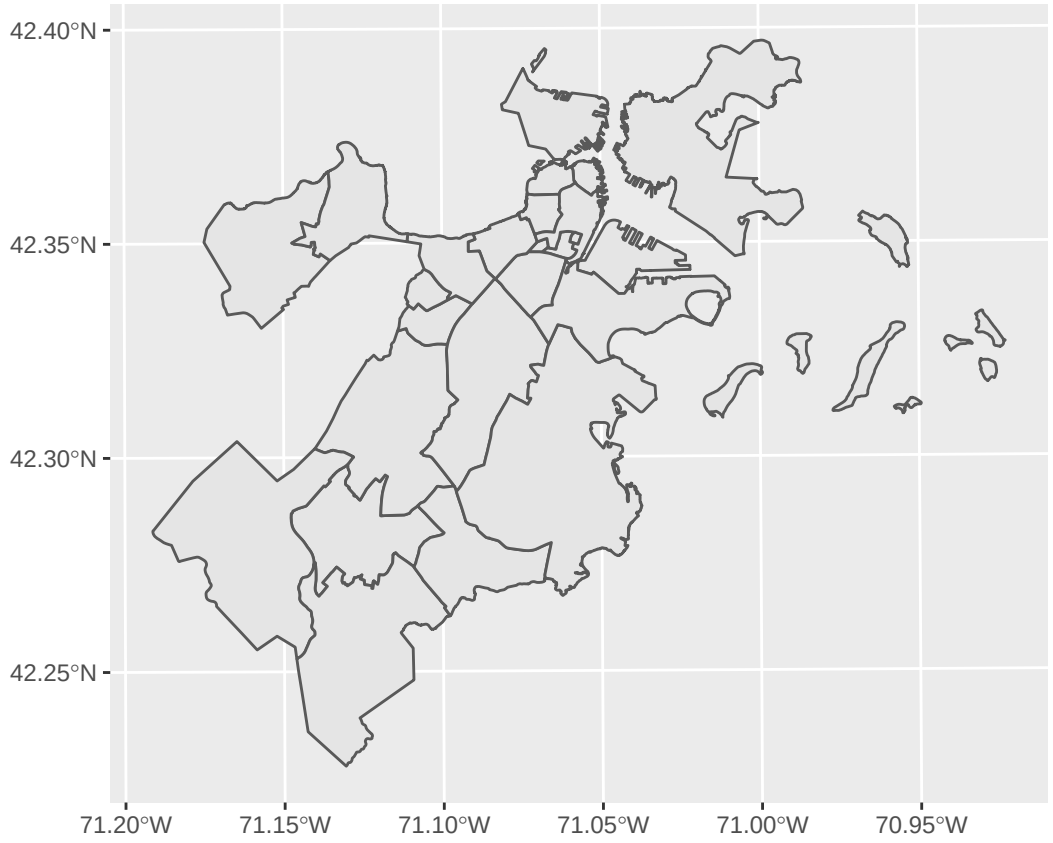
2/16/2022

In this assignment, I'd like to look at the distribution of blue bike stations across different neighborhoods of Boston. I'm interested to see which neighborhoods have the most stations, and if the quantity of stations is proportional to the population of that neighborhood.

First, let's read in the data set on the Boston neighborhoods.

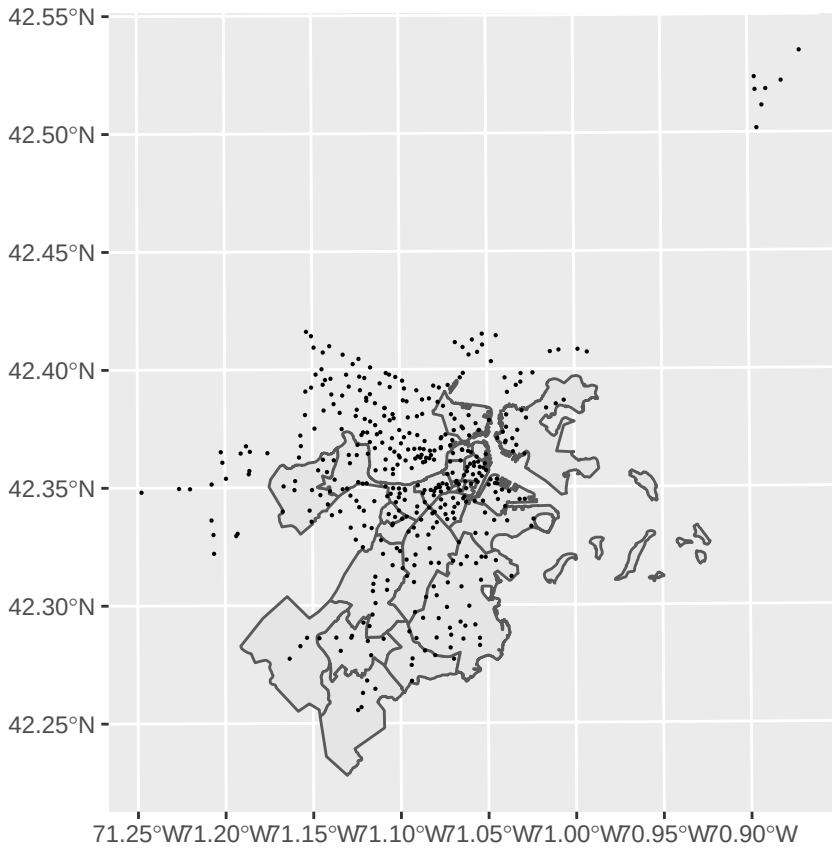
Before we do anything with blue bikes, let's plot the neghborhoods using ggplot to see what we're working with.

```
ggplot() + geom_sf(data = boston_neighborhoods)
```



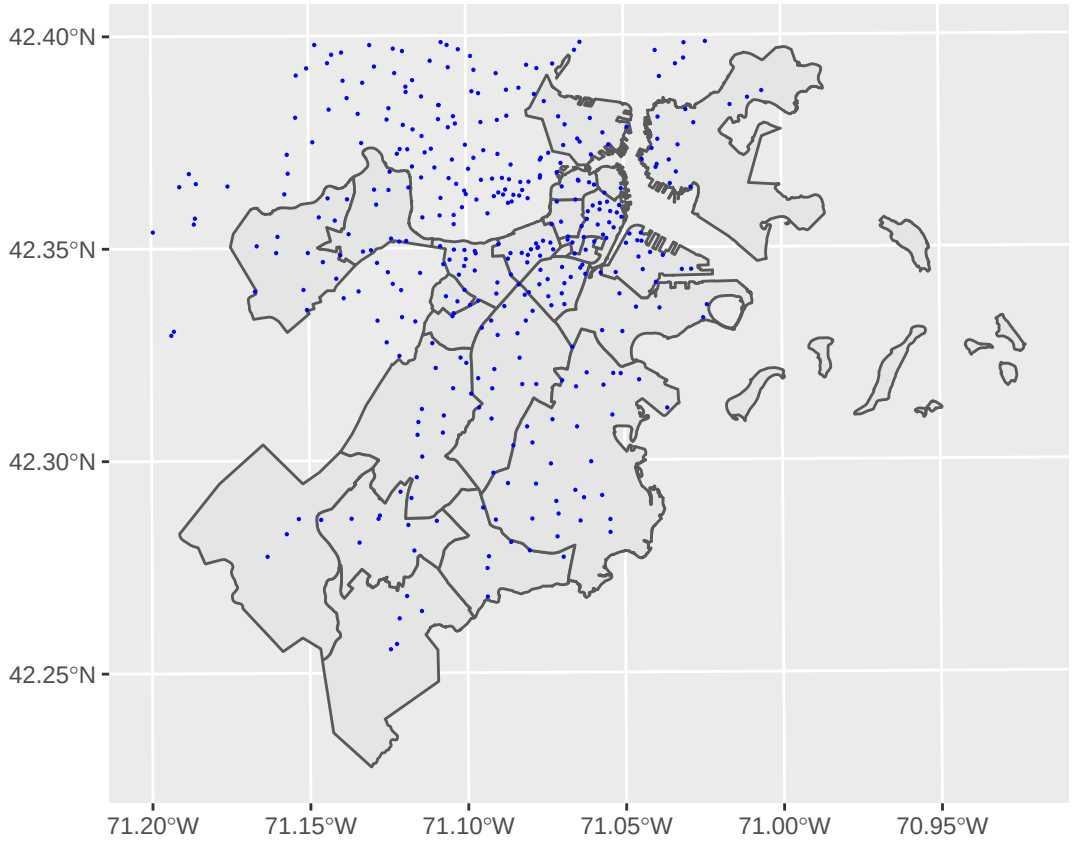
Now, let's read in the data set on blue bike stations as a shape file and overlay this data with the neighborhood data.

```
# Data source: https://data.boston.gov/dataset/blue-bike-stations1
blue_bikes <- read_sf("Blue_Bike_Stations/Blue_Bike_Stations.shp")

ggplot() + geom_sf(data = boston_neighborhoods) + geom_sf(data = blue_bikes, size = .1)
```



Let's clean up our data set a little bit to get a better look at what's going on here. We only need to see the blue bike data points that are within the bounds of the Boston neighborhoods.

```
#clean data to only have relevant points
blue_bikes <- subset(blue_bikes, Latitude < 42.4 & Longitude > -71.2)

#overlay points onto neighborhoods
ggplot() + geom_sf(data = boston_neighborhoods) + geom_sf(data = blue_bikes, size = .1, color = "blue")
```



The next step is to see which neighborhoods the points fall inside of. To do this, we can transform each of our data frames using st_transform. This will allow us to use sf's intersects function.

```
#transform sets
bikes_trans <- st_transform(blue_bikes, 2163)
neighborhood_trans <- st_transform(boston_neighborhoods, 2163)

#intersect data
intersections <- blue_bikes %>% mutate(
  neighborhood = as.integer(st_intersects(bikes_trans, neighborhood_trans)))

intersections
```

```
## Simple feature collection with 377 features and 9 fields
## Geometry type: POINT
## Dimension:     XY
## Bounding box:  xmin: -7925885 ymin: 5199344 xmax: -7904363 ymax: 5220853
## Projected CRS: WGS 84 / Pseudo-Mercator
## # A tibble: 377 x 10
##    Number Name           Latitude Longitude District Public_ Total_dock ObjectId
##  * <chr>  <chr>             <dbl>     <dbl> <chr>    <chr>        <int>    <int>
##  1 M32044 Huron Ave At ~     42.4     -71.1 Cambrid~ Yes             18        1
##  2 C32100 Hyde Park Ave~     42.3     -71.1 Boston   Yes              0        2
##  3 C32089 Hyde Park Ave~     42.3     -71.1 Boston   Yes             15        3
##  4 E32013 Hyde Park Lib~     42.3     -71.1 Boston   Yes             15        4
##  5 E32003 Hyde Square --     42.3     -71.1 Boston   Yes             15        5
```

```
##  6 C32020 ID Building E~     42.3     -71.0 Boston   Yes             17        6
##  7 C32021 ID Building W~     42.3     -71.0 Boston   Yes             19        7
##  8 C32025 Ink Block - H~     42.3     -71.1 Boston   Yes             15        8
##  9 M32062 Inman Square ~     42.4     -71.1 Cambrid~ Yes             25        9
## 10 A32011 Innovation La~     42.4     -71.1 Boston   Yes             19       10
## # ... with 367 more rows, and 2 more variables: geometry <POINT [m]>,
## #   neighborhood <int>
```

This data frame now contains all of the blue bike data, with a new added attribute called intersection, which tells us which neighborhood it falls into, or NA if it doesn't.

To see which neighborhoods the bike stations fall into, let's use the plyr library:

```
count(intersections, c("neighborhood"))
```

```
##    neighborhood freq
## 1             1    9
## 2             2   14
## 3             3    1
## 4             4    5
## 5             7    3
## 6             8    2
## 7             9   19
## 8            10   14
## 9            11   14
## 10           12   16
## 11           13   11
## 12           14    6
## 13           15    4
## 14           16   19
## 15           17   14
## 16           18   12
## 17           19    3
## 18           20    5
## 19           21    7
## 20           22   28
## 21           23   13
## 22           24    9
## 23           25   11
## 24           NA  138
```

From this, we can see that neighborhood 22 has the most blue bike stations with 28.

If we want to observe the mean, median and other statistics on each station's neighborhood, we can use the summary method as shown next.

```
summary(as.numeric(intersections$neighborhood))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00   10.00   16.00   14.68   22.00   25.00     138
```

Note to grader: If I had more time, I would visualize this so we could see what the neighborhoods actually are instead of just saying district x. I had quite a bit of trouble figuring out syntax for this assignment but that would have been my next step.