2015 NYC Street Tree Census EDA by Sayed Malawi

This report contains an exploration of data from the 2015 New York City TreesCount street tree census, organized by the city's Department of Parks & Recreation. Data on over 680,000 of the city's street trees (i.e. not including parks, etc.) was collected by staff and volunteers. Each observation contains up to 40 pieces of information, such as species, health, and size, as well as highly accurate location data. As a student worker at my college's arboretum, this dataset was of interest to me; additionally, I was inspired by the use of map plots in some of the example projects.

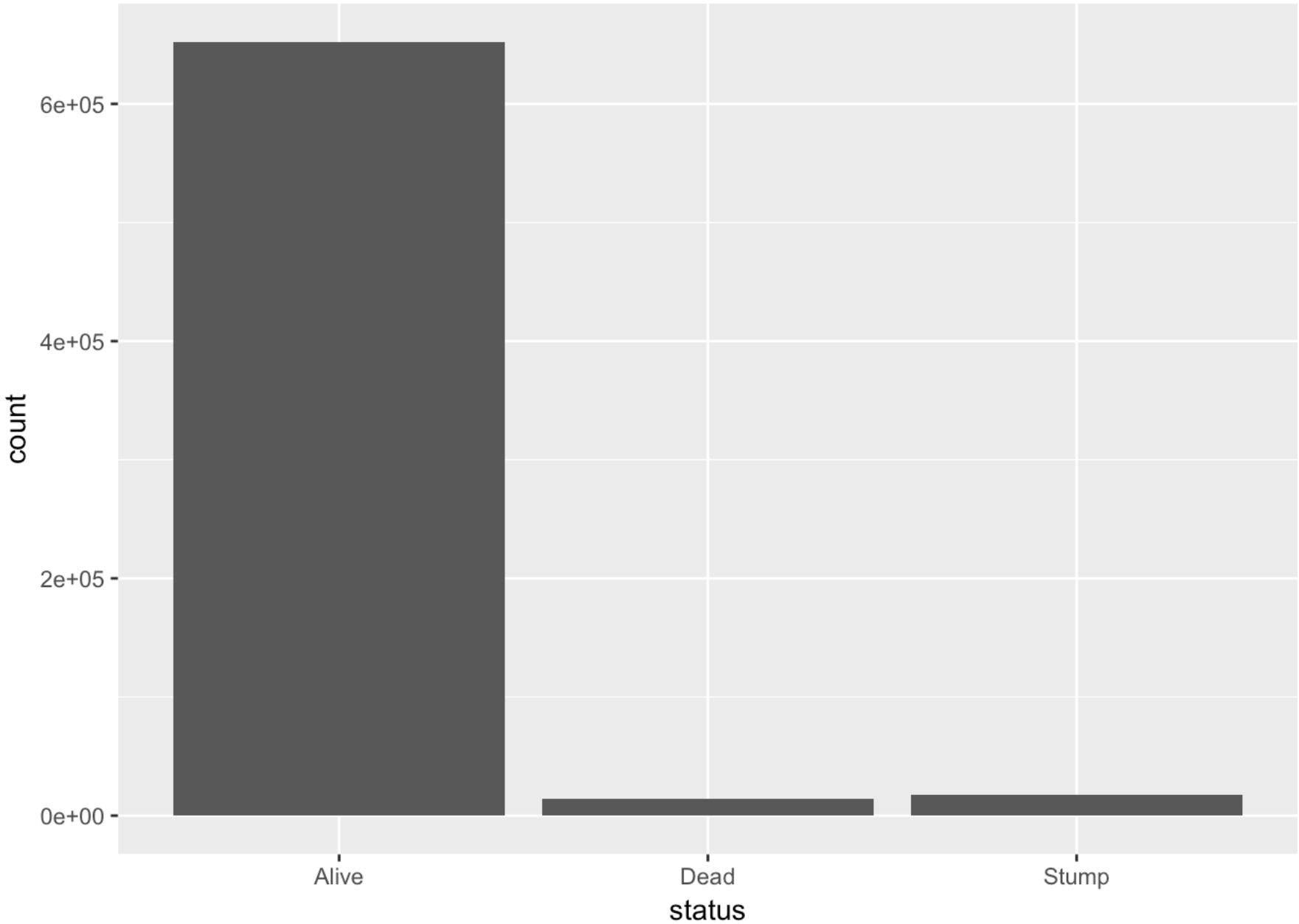
Univariate Plots Section

```
## 'data.frame':    683788 obs. of  41 variables:
##  $ tree_id   : int  180683 200540 204026 204337 189565 190422 190426 208649 209610
192755 ...
##  $ block_id  : int  348711 315986 218365 217969 223043 106099 106099 103940 407443
207508 ...
##  $ created_at: Factor w/ 483 levels "01/01/2016","01/02/2016",..: 320 334 338 338
326 326 326 342 344 328 ...
##  $ tree_dbh  : int  3 21 3 10 21 11 11 9 6 21 ...
##  $ stump_diam: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ curb_loc  : Factor w/ 2 levels "OffsetFromCurb",..: 2 2 2 2 2 2 2 2 2 1 ...
##  $ status    : Factor w/ 3 levels "Alive","Dead",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ health    : Factor w/ 4 levels "","Fair","Good",..: 2 2 3 3 3 3 3 3 3 2 ...
##  $ spc_latin : Factor w/ 133 levels "","Acer","Acer buergerianum",..: 13 110 59 59
126 59 59 126 59 93 ...
##  $ spc_common: Factor w/ 133 levels "","'Schubert' chokecherry",..: 98 91 63 63 8
63 63 8 63 75 ...
##  $ steward   : Factor w/ 5 levels "","1or2","3or4",..: 5 5 2 5 5 2 2 5 5 5 ...
##  $ guards    : Factor w/ 5 levels "","Harmful","Helpful",..: 4 4 4 4 4 3 3 4 4 4 .
..
##  $ sidewalk  : Factor w/ 3 levels "","Damage","NoDamage": 3 2 2 2 2 3 3 3 3 3 ...
##  $ user_type : Factor w/ 3 levels "NYC Parks Staff",..: 2 2 3 3 3 3 3 3 2 2 ...
##  $ problems  : Factor w/ 233 levels "","BranchLights",..: 48 85 48 85 85 48 48 7 4
8 48 ...
##  $ root_stone: Factor w/ 2 levels "No","Yes": 1 2 1 2 2 1 1 1 1 1 ...
##  $ root_grate: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 2 1 1 ...
##  $ root_other: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ trunk_wire: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ trnk_light: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ trnk_other: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ brch_light: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ brch_shoe : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ brch_other: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ address   : Factor w/ 408701 levels "0 5 AVENUE","1",..: 15840 80258 253362 647
0 317912 363641 38349 221108 329842 326509 ...
```

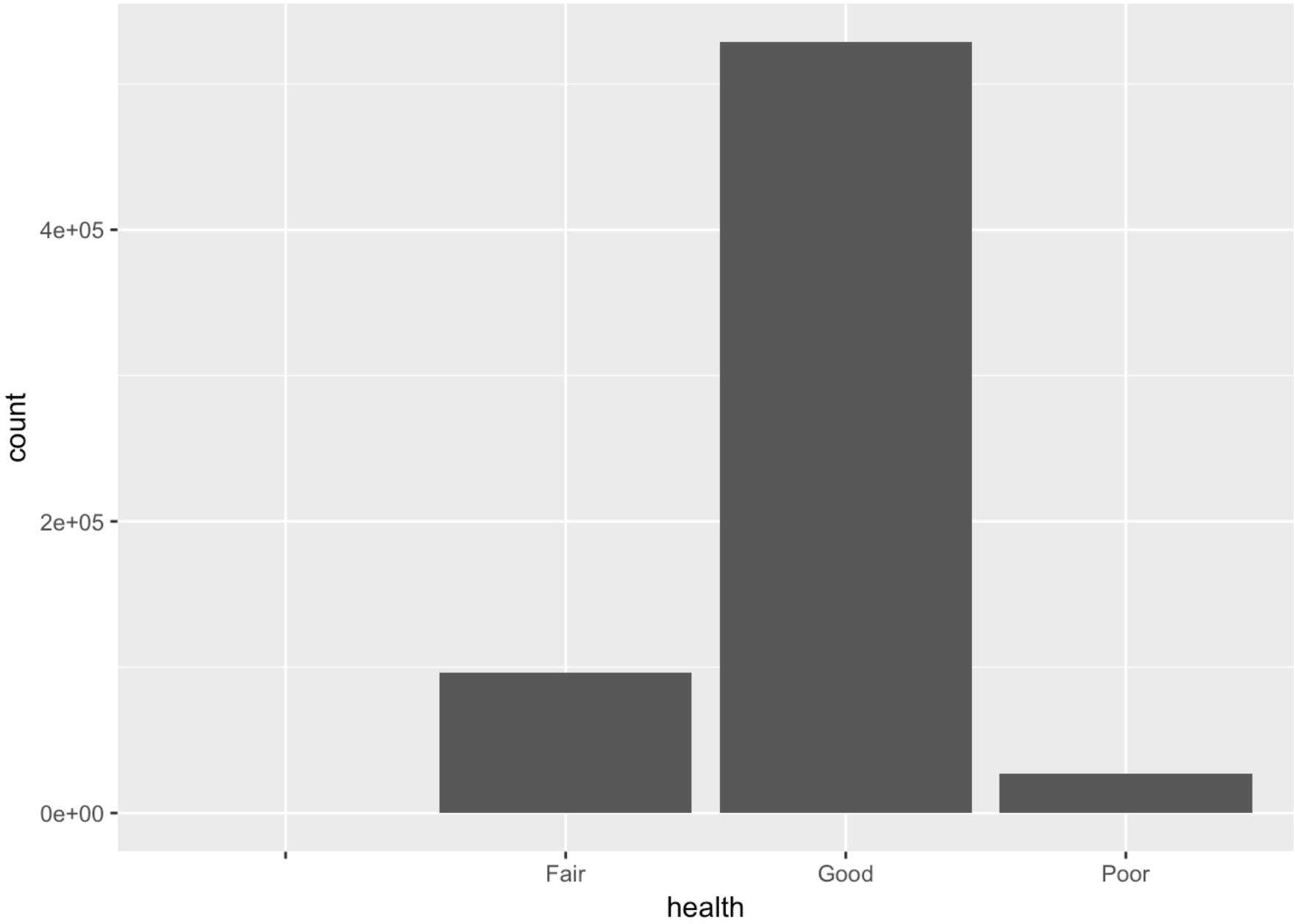
```
##  $ zipcode   : int  11375 11357 11211 11211 11215 10023 10023 10019 10305 11223 ..
.
##  $ zip_city  : Factor w/ 48 levels "Arverne","Astoria",..: 17 46 7 7 7 31 31 31 44
7 ...
##  $ cb_num    : int  406 407 301 301 306 107 107 104 502 313 ...
##  $ borocode  : int  4 4 3 3 3 1 1 1 5 3 ...
##  $ boroname  : Factor w/ 5 levels "Bronx","Brooklyn",..: 4 4 2 2 2 3 3 3 5 2 ...
##  $ cncldist  : int  29 19 34 34 39 3 3 3 50 47 ...
##  $ st_assem  : int  28 27 50 53 44 67 67 75 64 45 ...
##  $ st_senate : int  16 11 18 18 21 27 27 27 23 23 ...
##  $ nta       : Factor w/ 188 levels "BK09","BK17",..: 125 152 46 46 17 95 95 96 17
7 7 ...
##  $ nta_name  : Factor w/ 188 levels "Allerton-Pelham Gardens",..: 66 181 56 56 129
96 96 34 74 75 ...
##  $ boro_ct   : int  4073900 4097300 3044900 3044900 3016500 1014500 1014500 101270
0 5006400 3037402 ...
##  $ state     : Factor w/ 1 level "New York": 1 1 1 1 1 1 1 1 1 1 ...
##  $ latitude  : num  40.7 40.8 40.7 40.7 40.7 ...
##  $ longitude : num  -73.8 -73.8 -73.9 -73.9 -74 ...
##  $ x_sp      : num  1027431 1034456 1001823 1002420 990914 ...
##  $ y_sp      : num  202757 228645 200717 199244 182202 ...
```

There are a lot of factor variables to work with! Many are related to location and plant bed problems.

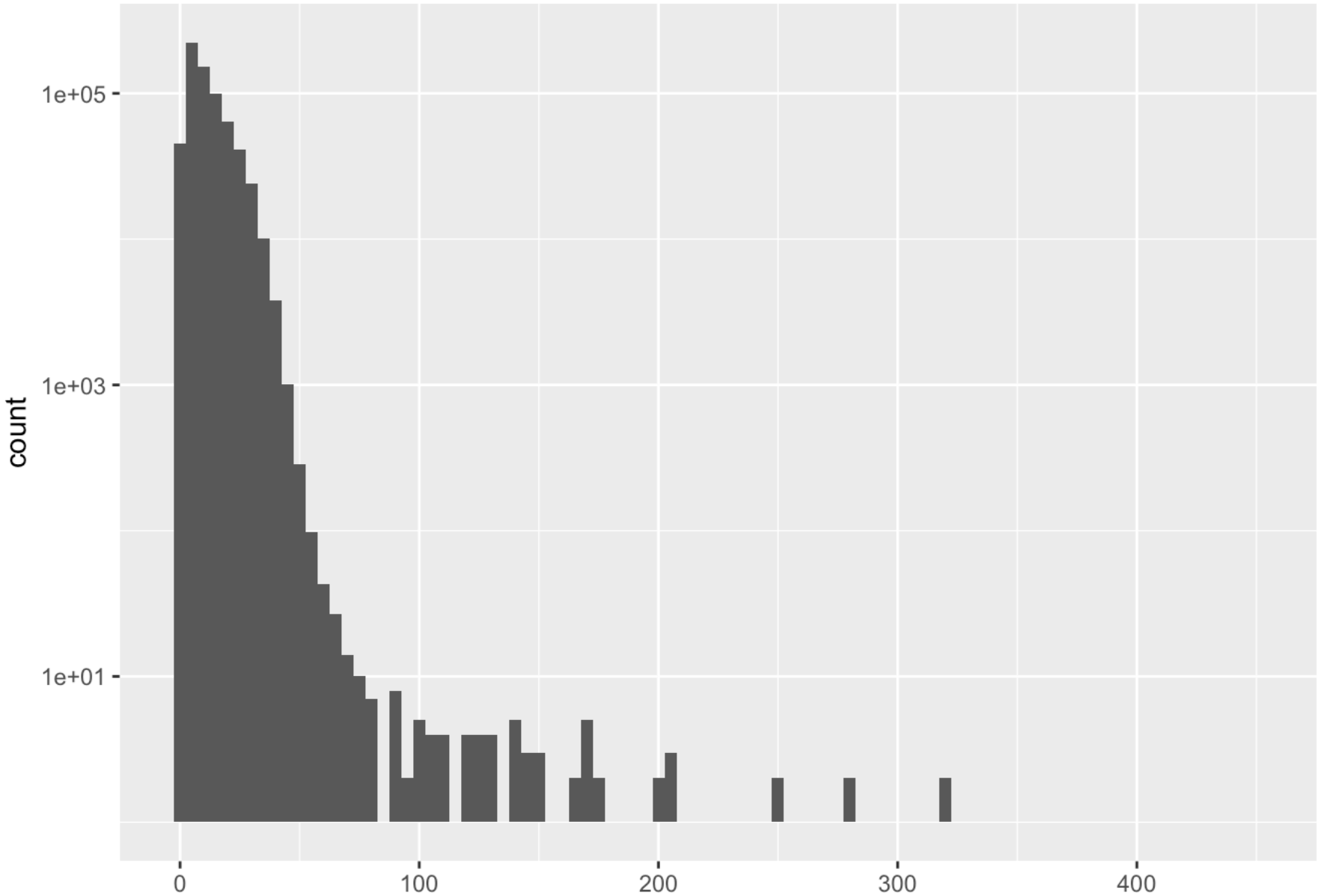
We can start with some simple plots investigating one variable from the dataset.



Most of the street trees happen to be alive, thankfully. (“Dead” refers to still-standing dead trees, as opposed to stumps.)



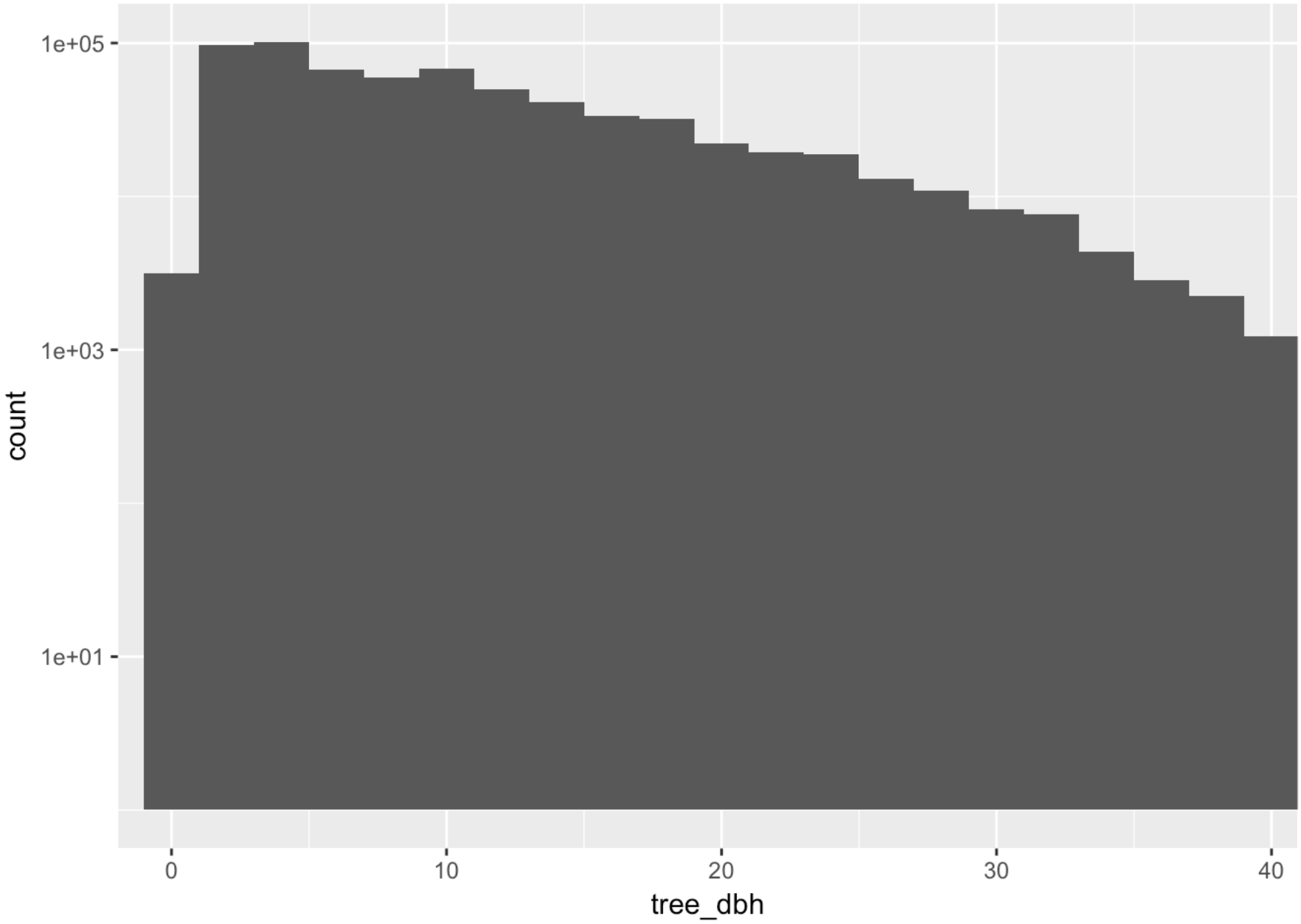
Of the trees that are alive, it looks like about 70-80% are in good health.



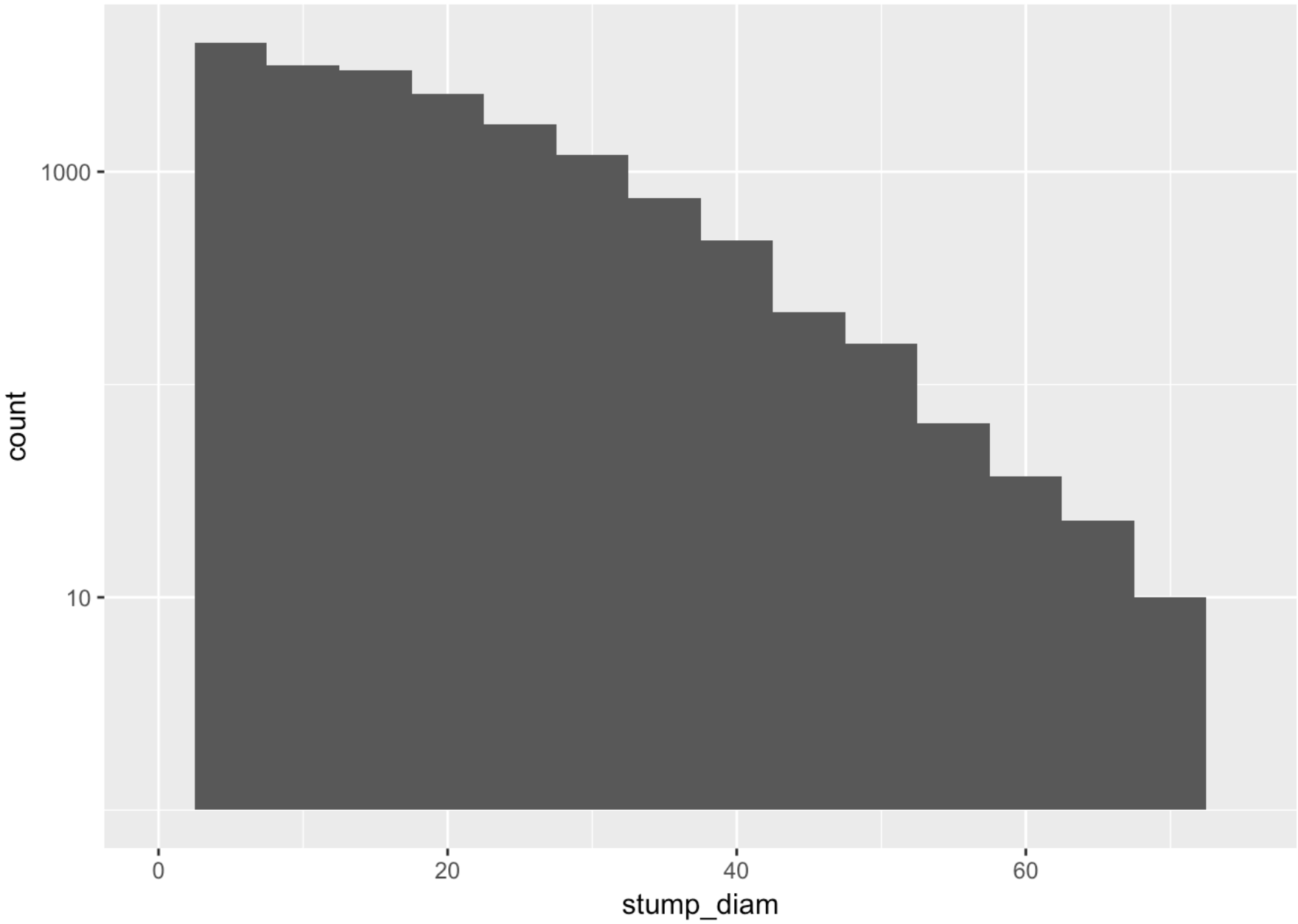
This plot shows the diameters of all non-stump trees, measured in inches, with a log scale applied to the y axis. Evidently, the vast majority of street trees are on the small or medium side, with diameters in the 0-50 inch range.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    5.00   10.00   11.58   16.00  450.00
```

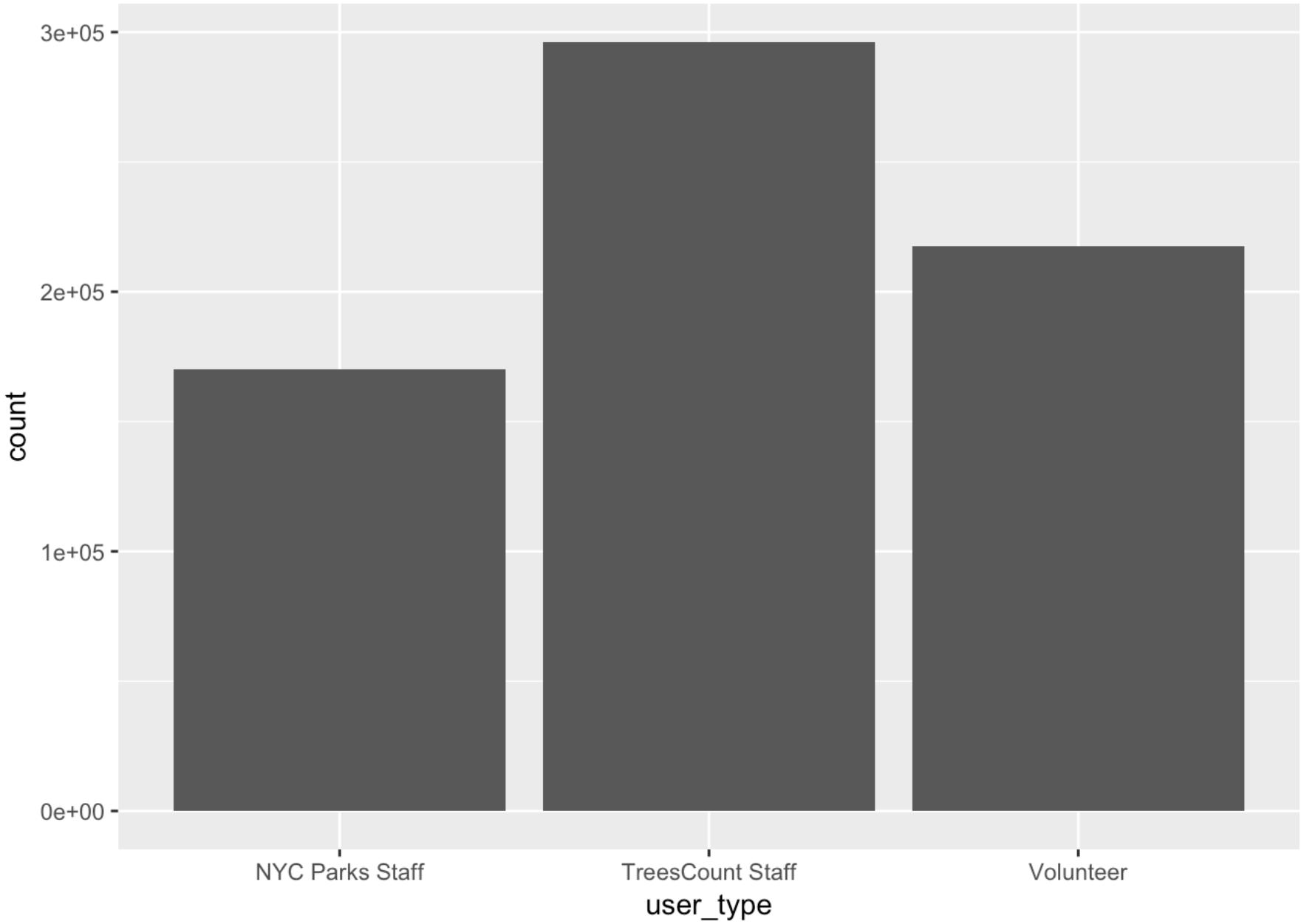
Some of the street trees with diameters in the hundred-inch range may be victims of data misentry. Plotting without the outliers may give a better visualization of the distribution.



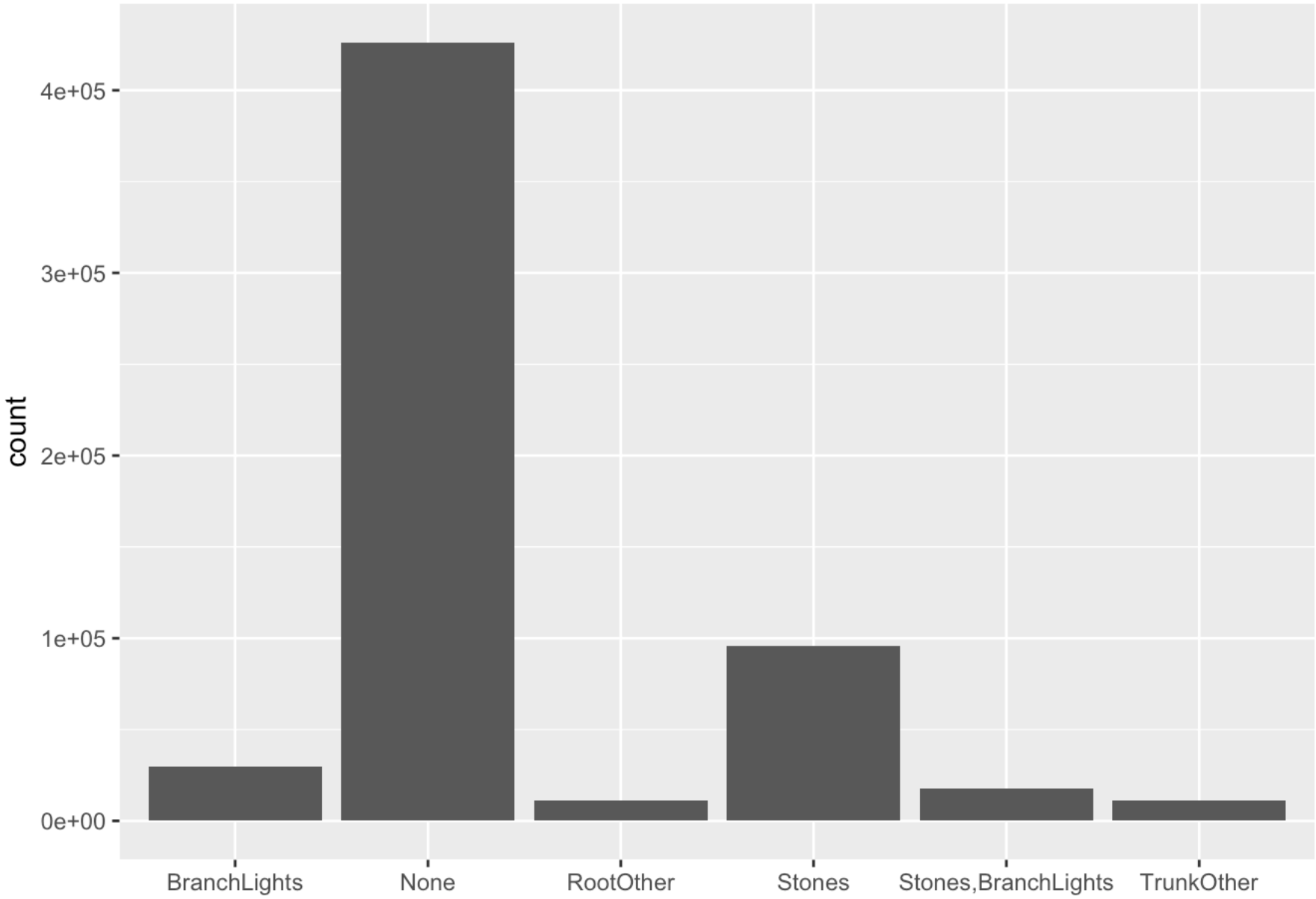
As the above graph shows, the distribution is still right-skewed, but not as severely as previously thought.



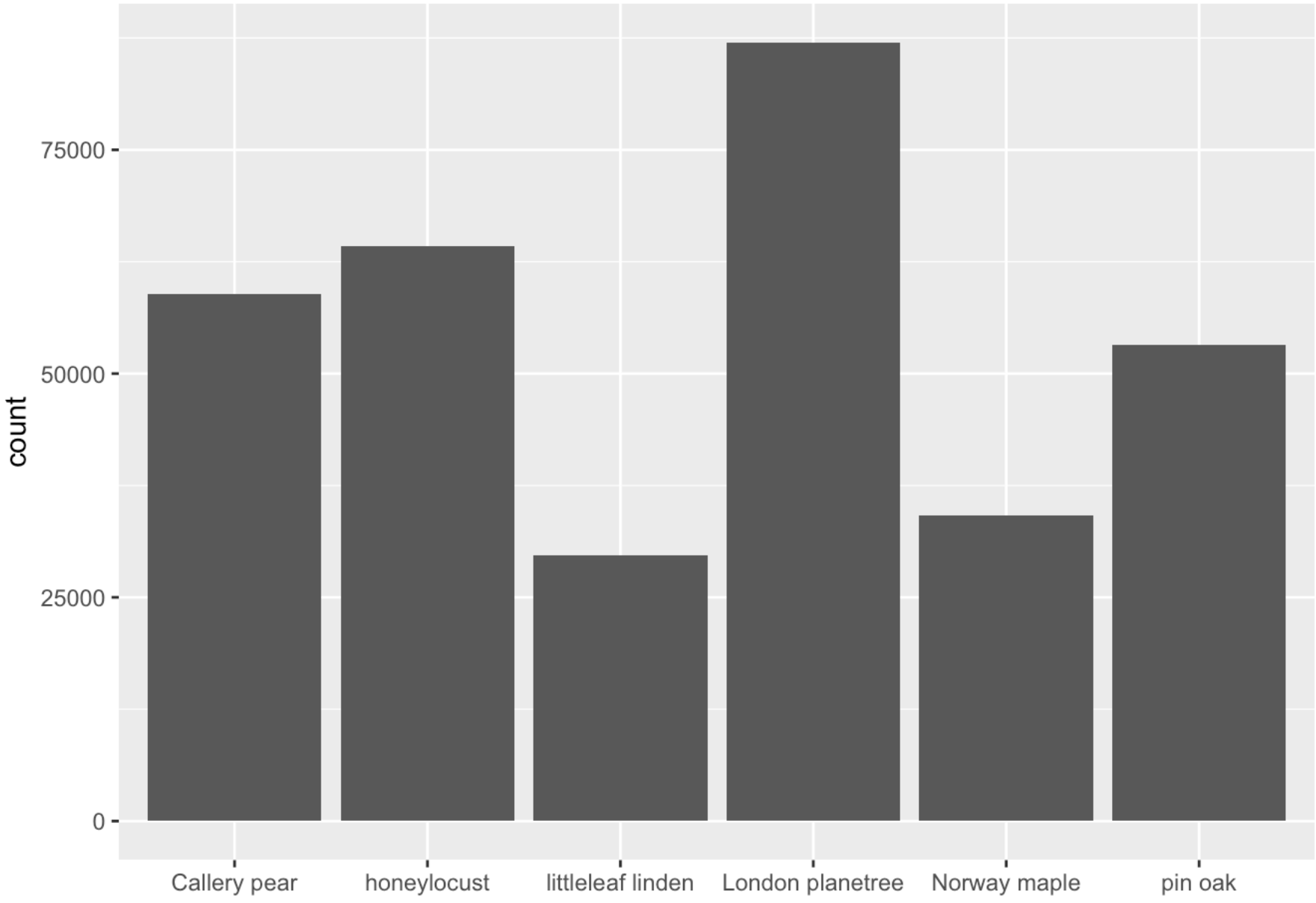
Stump sizes exhibit a similar distribution.



Here, we can see the number of trees cataloged by Parks Department staff, census staff, and volunteers. Significant contributions were made by each of the three groups. It might be interesting to see if volunteers' ratings of trees differed at all from those of staff.



Looking at the six most common values in the ‘problems’ field, it appears that the majority of trees had no reported problems, with ‘Stones’ and ‘BranchLights’ being the most significant contributors otherwise. ‘Stones’ corresponds to the presence of paving stones in the tree bed, which may cause root problems. ‘BranchLights’ refers to the presence of harmful light installations on the trunk of the tree.



This barplot shows the counts of the six most common species of New York City street trees. The most prevalent species are the London planetree, honey locust, Callery pear, and pin oak; with over 50,000 specimens each, these four species account for over 1/3 of the trees reported.

Univariate Analysis

What is the structure of your dataset?

The dataset contains information for 683,788 street trees in New York City. Live and dead trees are included. 40 variables are included, mostly categorical or location-related.

What is/are the main feature(s) of interest in your dataset?

The most compelling part of this data is the potential to link the variables analyzed above to location information - for example, species or health distributions.

What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

The location data should be possible to combine with a city shapefile to create interesting map plots.

Did you create any new variables from existing variables in the dataset?

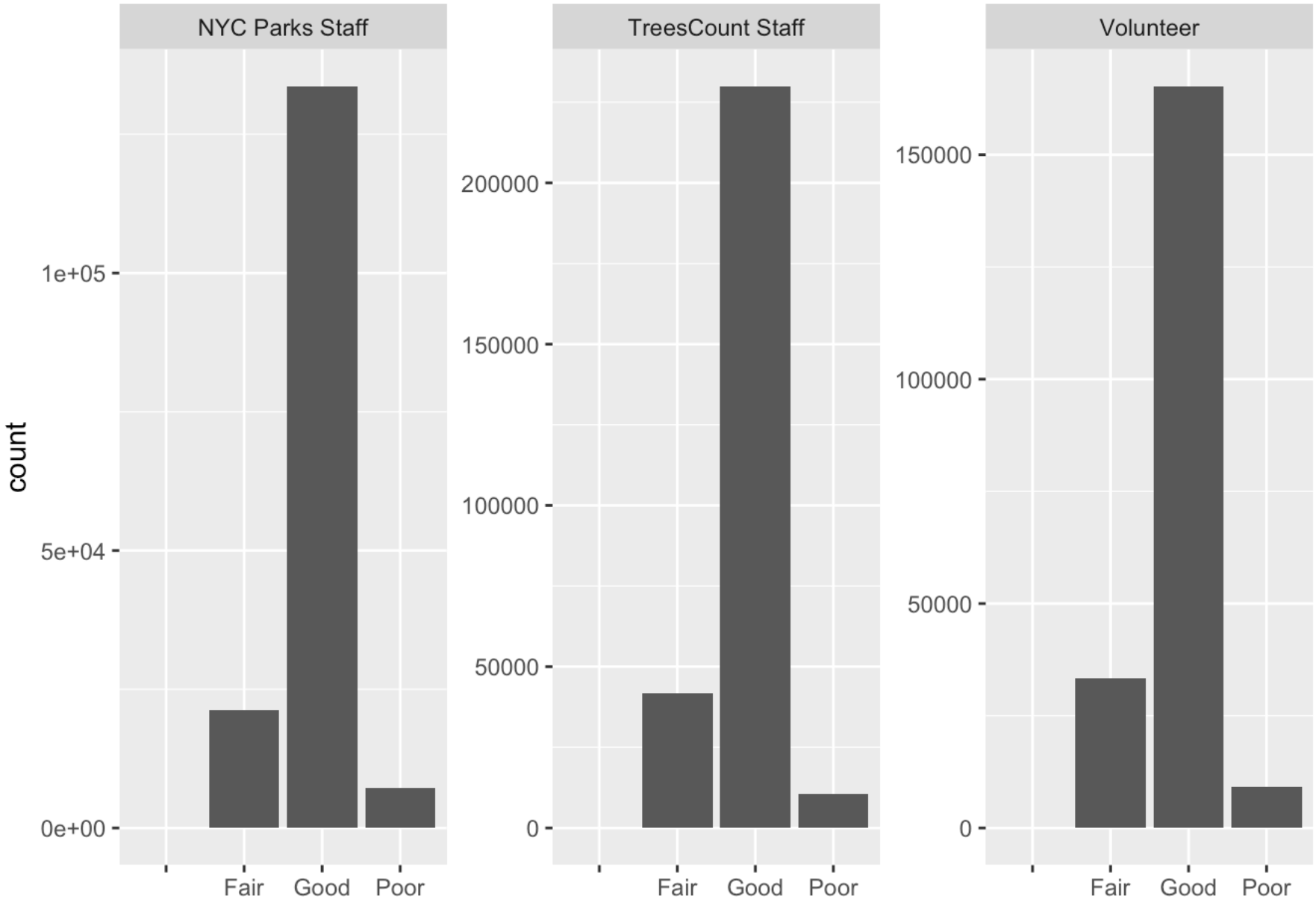
Not for univariate investigaion.

Of the features you investigated, were there any unusual distributions?
Did you perform any operations on the data to tidy, adjust, or change the form
of the data? If so, why did you do this?

The only numerical variables were tree and stump diameter, which exhibited relatively unsurprising distributions.

Bivariate Plots Section

One of the questions that arose from the univariate exploration was whether staff and volunteers exhibited differing behavior during evaluation of a tree's status. This can be investigated using faceted bar plots.

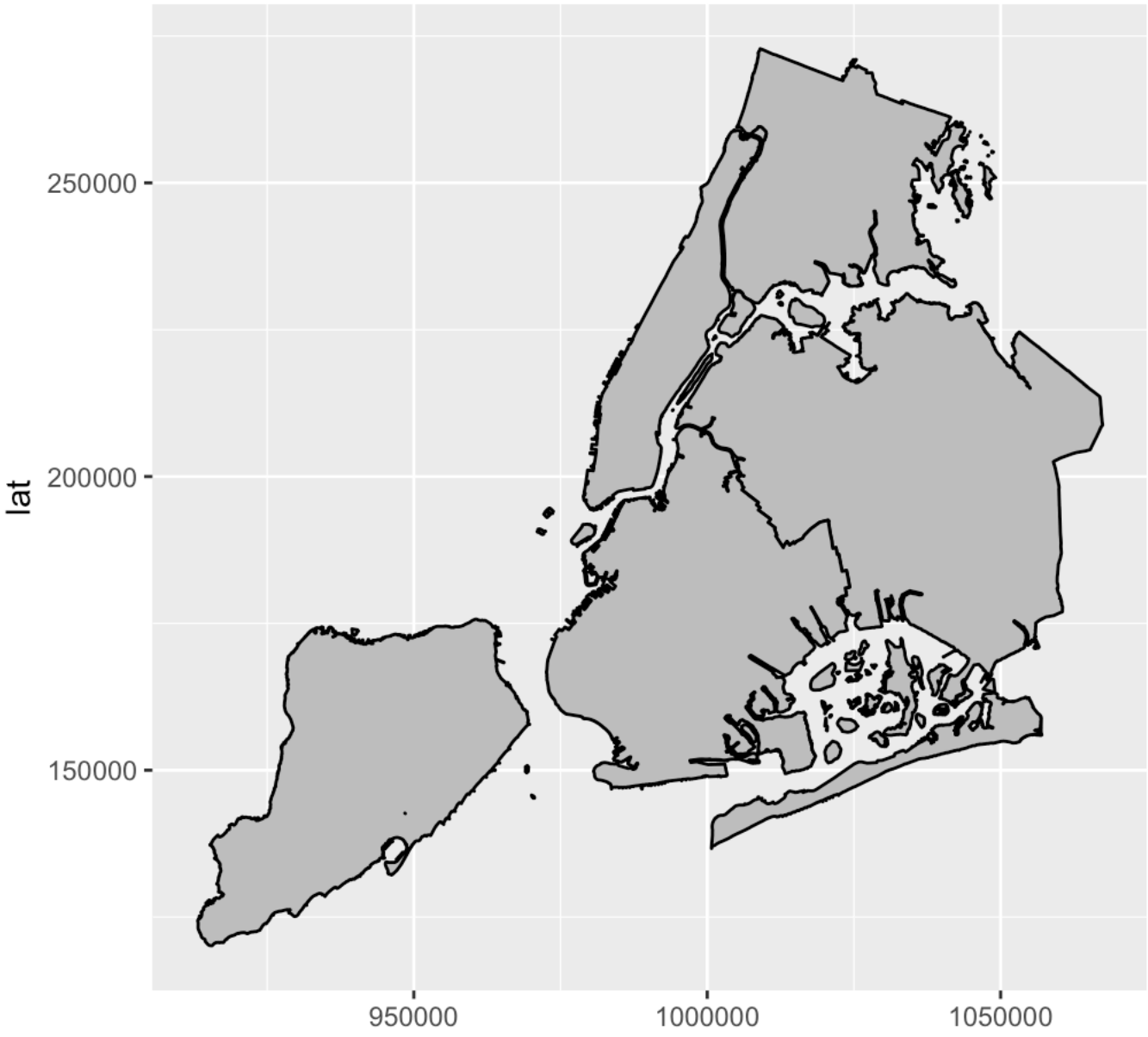


The distributions seem quite similar - it looks like volunteers were equally competent to staff in making assessments of tree health.

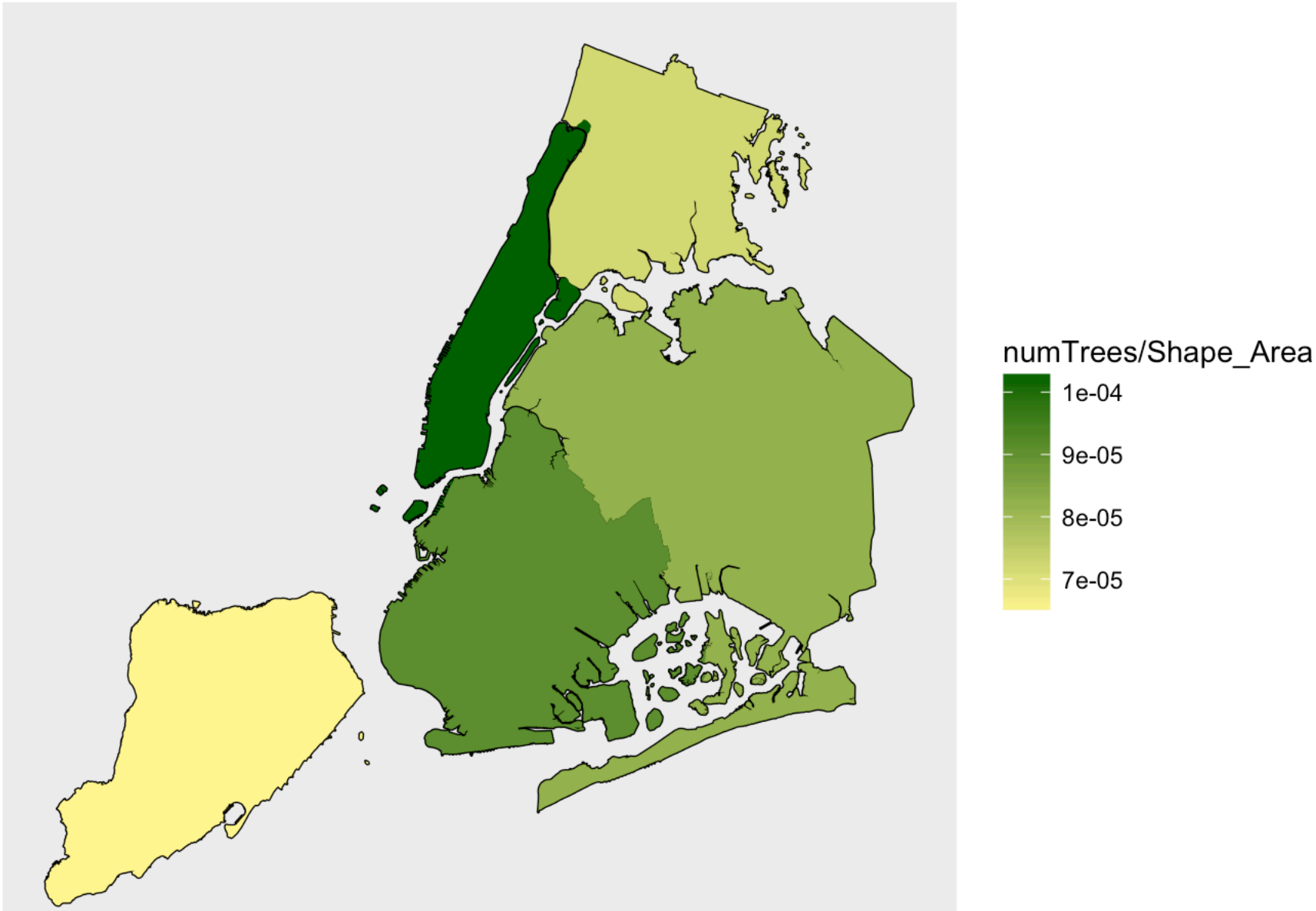
Let's start looking at the location data. The traditional way of creating map plots in R is via the maps package. The package includes maps for countries and states, but not cities, so for our purposes we need to find an appropriate shapefile and load it (preferably into a dataframe). This helpful post (https://github.com/tidyverse/ggplot2/wiki/plotting-polygon-shapefiles) was consulted to learn how to do just that using the rgdal and maptools packages. A shapefile for the five city boroughs can be obtained at the NYC Department of City Planning webpage (https://www1.nyc.gov/site/planning/data-maps/open-data/districts-download-metadata.page).

```
## OGR data source with driver: ESRI Shapefile
## Source: "Shapefiles/nybb.shp", layer: "nybb"
## with 5 features
## It has 4 fields
```

With the polygons laded, this post (http://eriqande.github.io/rep-res-web/lectures/making-maps-with-R.html) was used to better understand the plotting syntax for map shapes. Using geom_polygon(), a simple map of New York City can be generated.



By matching the borough names with those in the trees dataframe, we can use map plots to visualize our variables with respect to location. For example, the following plot illustrates the trees per area in each borough.

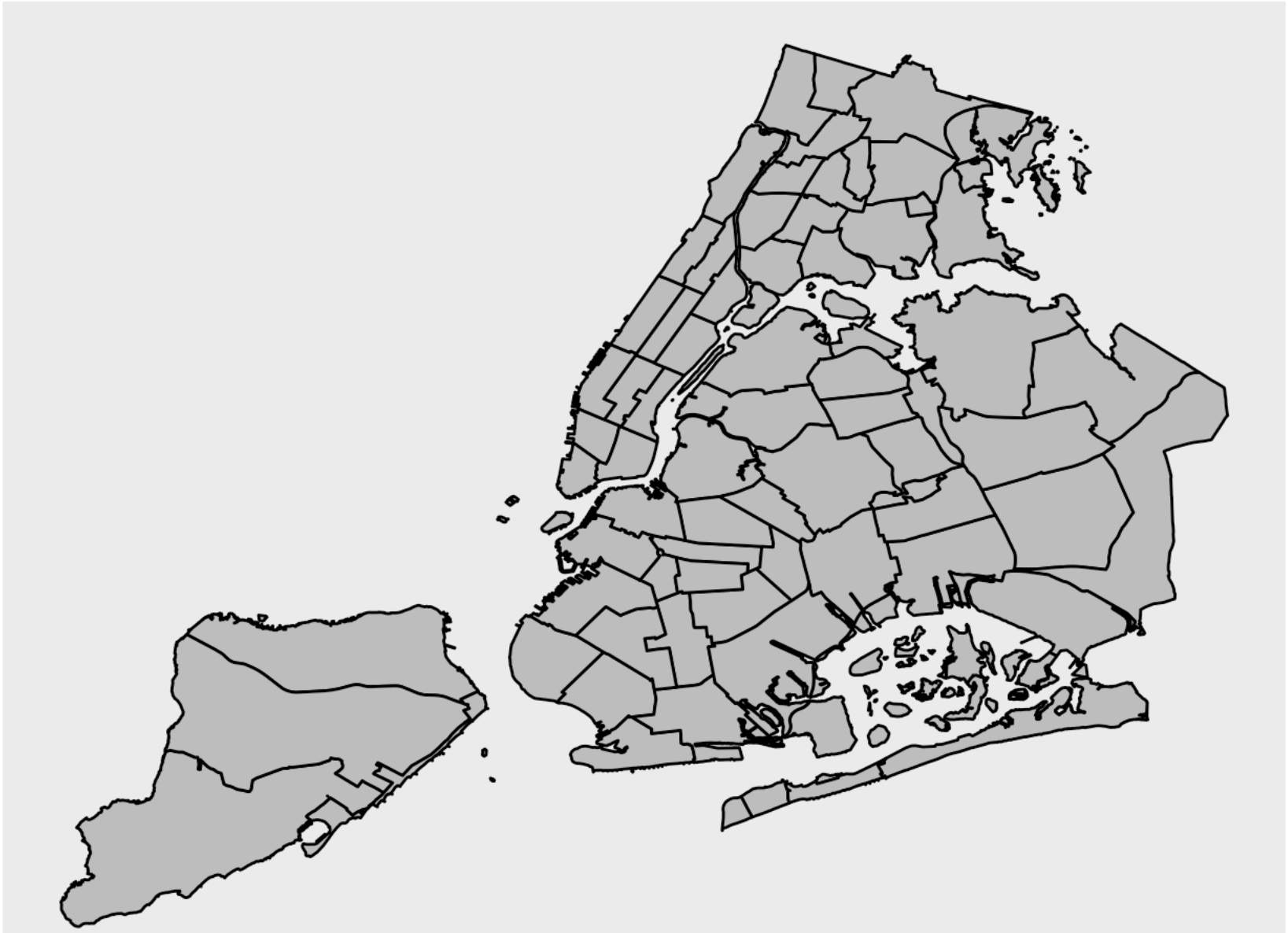


Judging by the legend, Manhattan has the greenest streets of the five boroughs, with about 50% more trees per area than Staten Island. Given that this dataset omits trees in parks and the like, it's a bit surprising that the most densely populated borough also exhibits the highest street tree density.

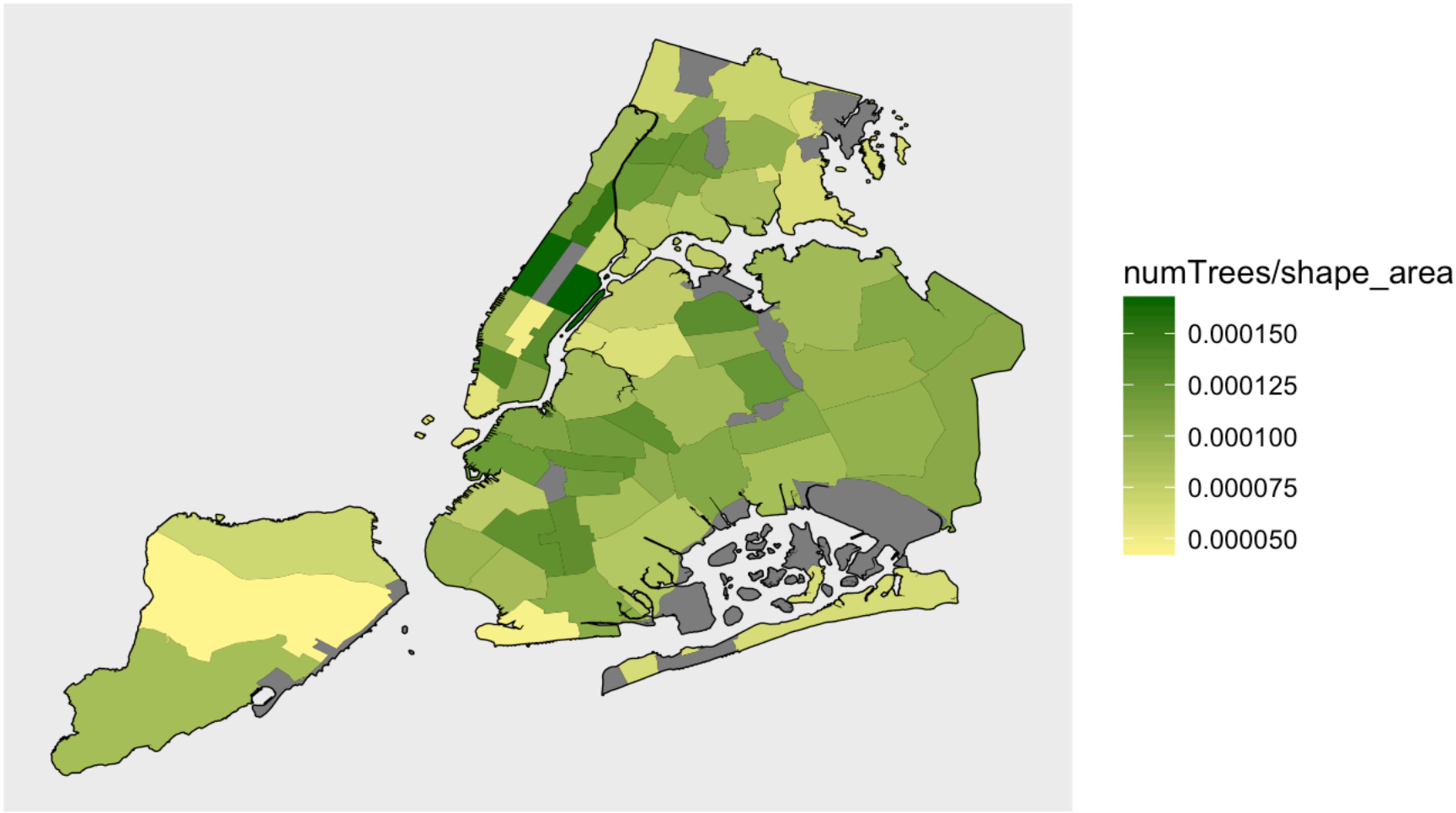
Although borough visualizations are definitely of interest, the information they contain could just as easily be gleaned from a table. It would be more interesting to divide the map into further subsections and then explore further variables. This can be done using a shapefile of the city's 59 community districts.

Loading and preparing such a shapefile was completed in a similar fashion to above, with the exception of obtaining the file from a NYC OpenData map (https://data.cityofnewyork.us/City-Government/Community-Districts/yfnk-k7r4/data), since the file from the City Planning website had some corrupt polygons.

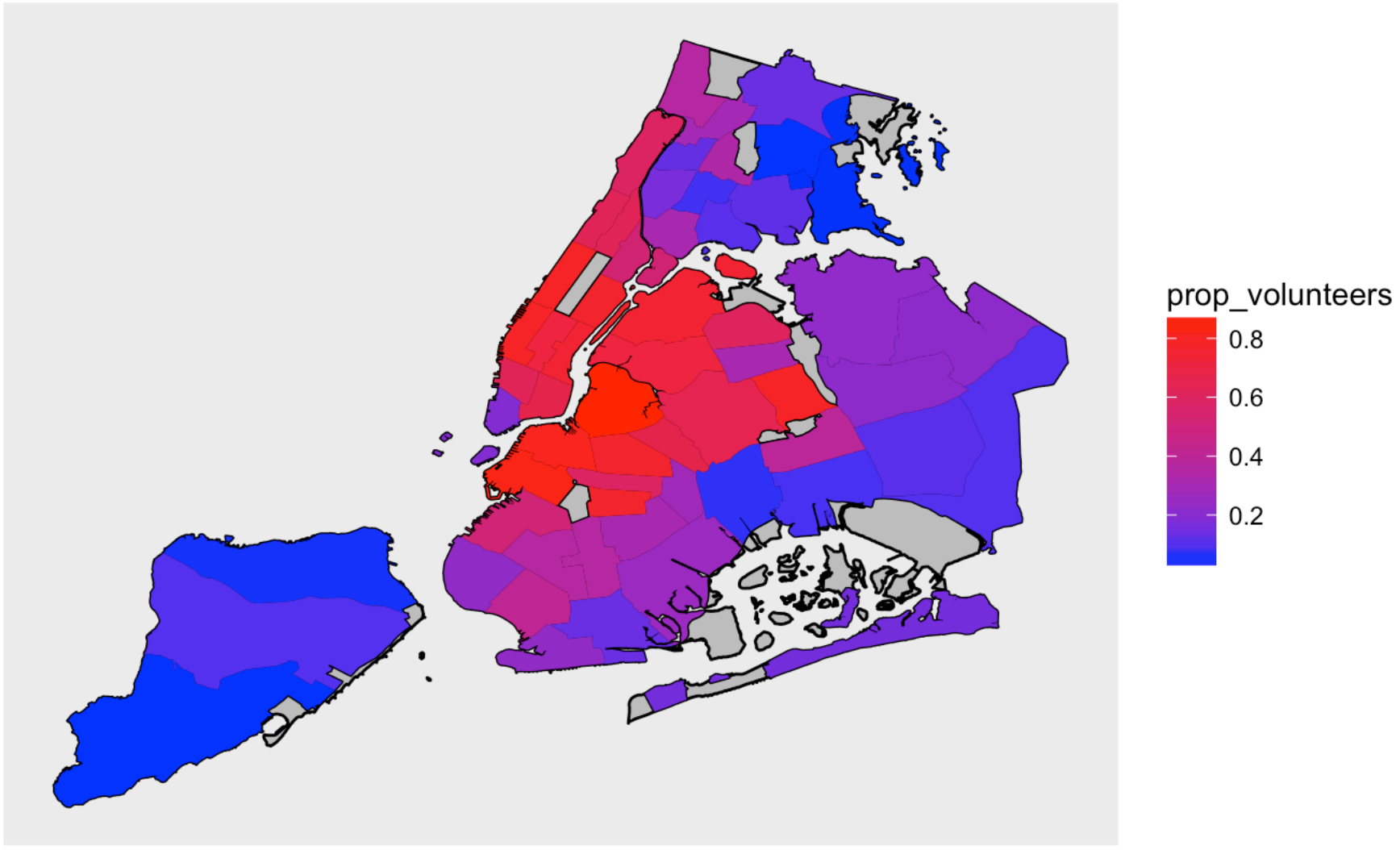
```
## OGR data source with driver: ESRI Shapefile
## Source: "Shapefiles/nycd.shp", layer: "nycd"
## with 71 features
## It has 3 fields
```



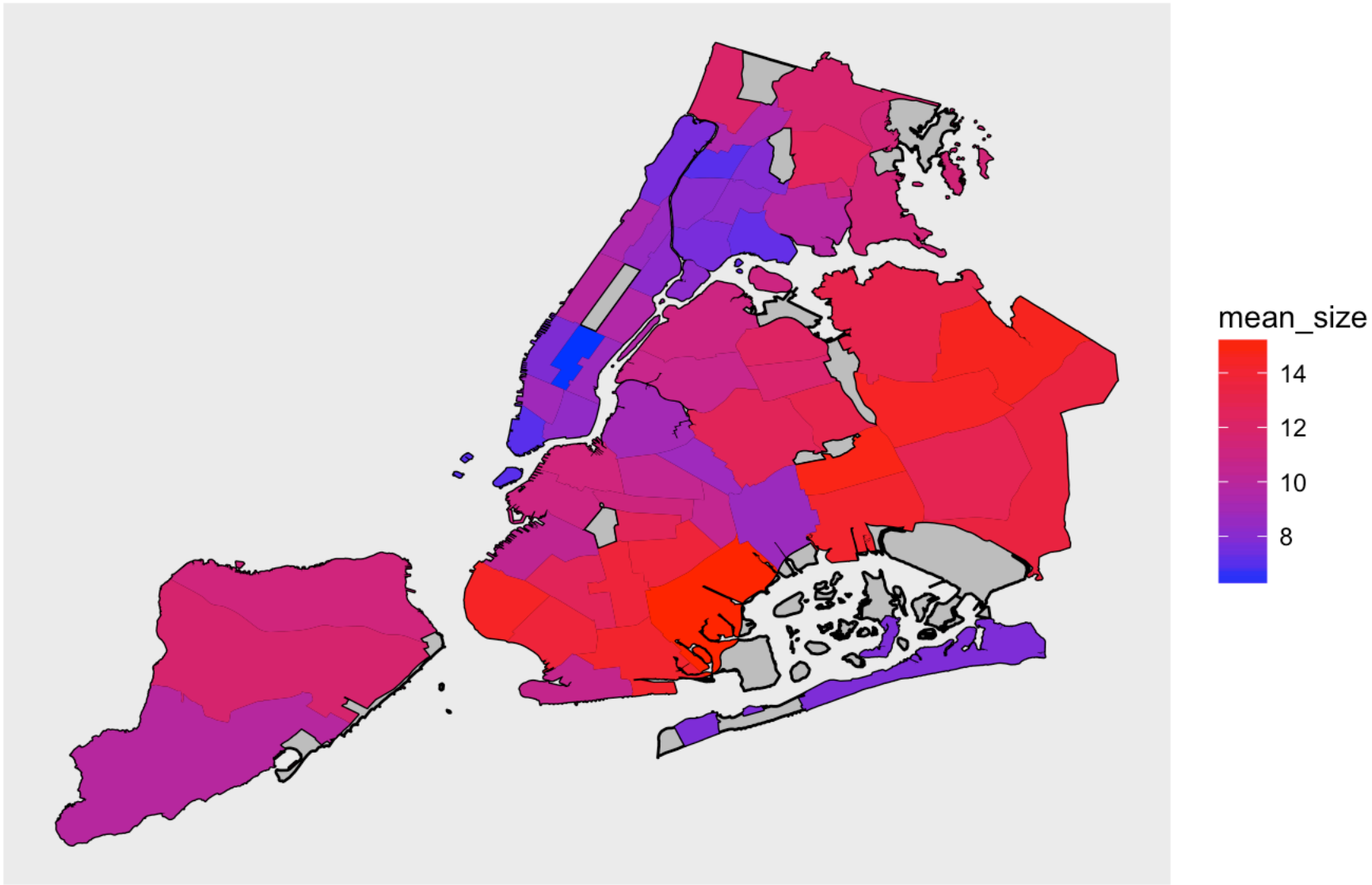
This map will allow for the display of more detailed information. Let’s see what the tree density looks like using this map.



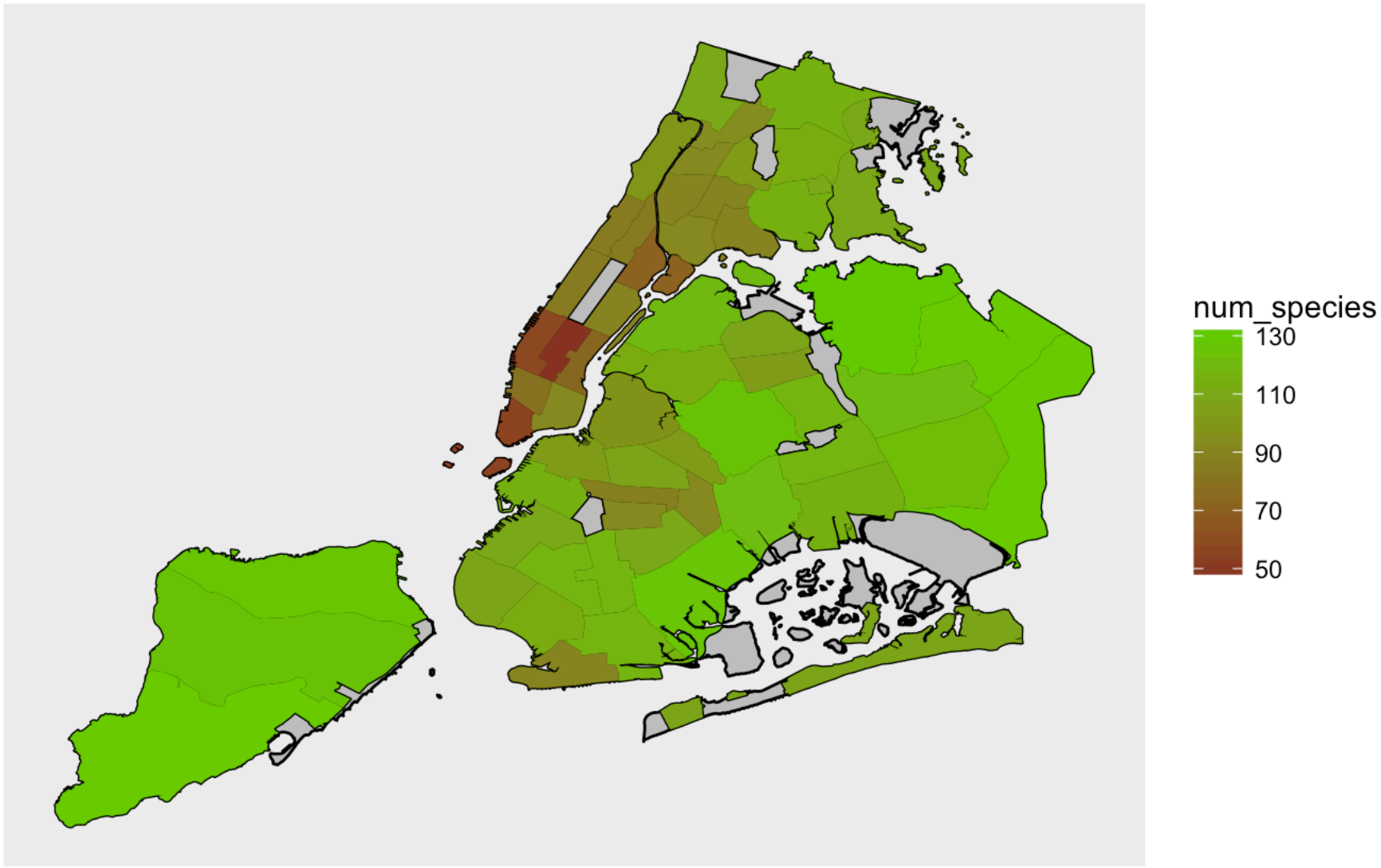
This shows that the areas of greatest tree density tend to be towards the center of the city, in areas such as Manhattan and much of Brooklyn.



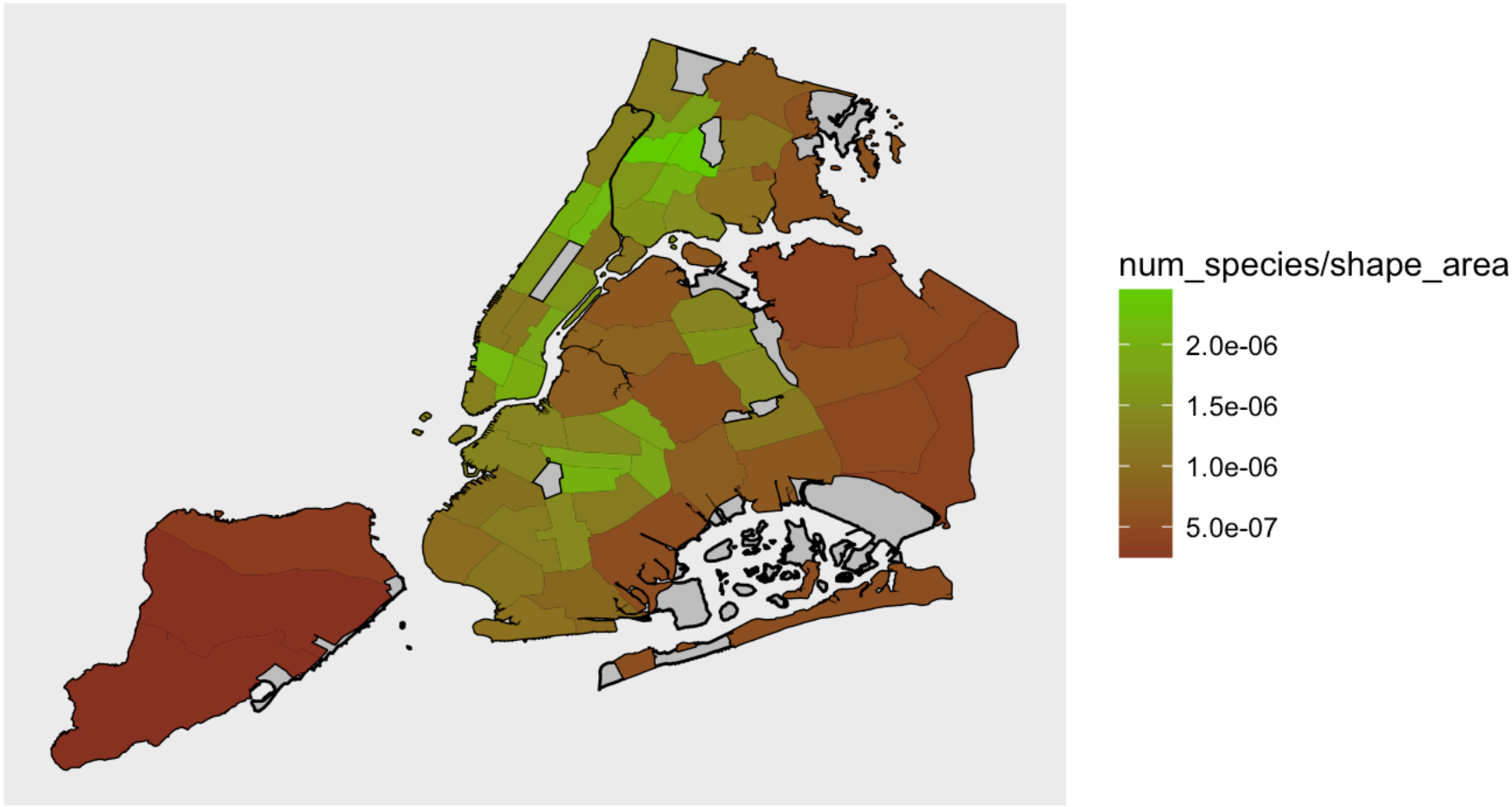
This plot illustrates the proportion of trees cataloged by volunteers (as opposed to staff) in each community district. Volunteer activity seems to mirror tree density, being higher in the central districts of the city. The legend shows that these differences are quite large, with volunteers cataloging around 10% of the trees in Staten Island as opposed to around 80% in north Brooklyn. Volunteer engagement is much stronger in the areas with higher tree density, even though tree density variation is not that large, excepting central Manhattan. The next tree census may benefit from additional promotion in the city's outer boroughs.



Here, the average tree diameter is displayed for each community district. The average street tree in much of Queens and Brooklyn is over a foot in diameter, while Manhattan and West Bronx street trees average significantly smaller; this makes sense given the densely populated developments of those neighborhoods.

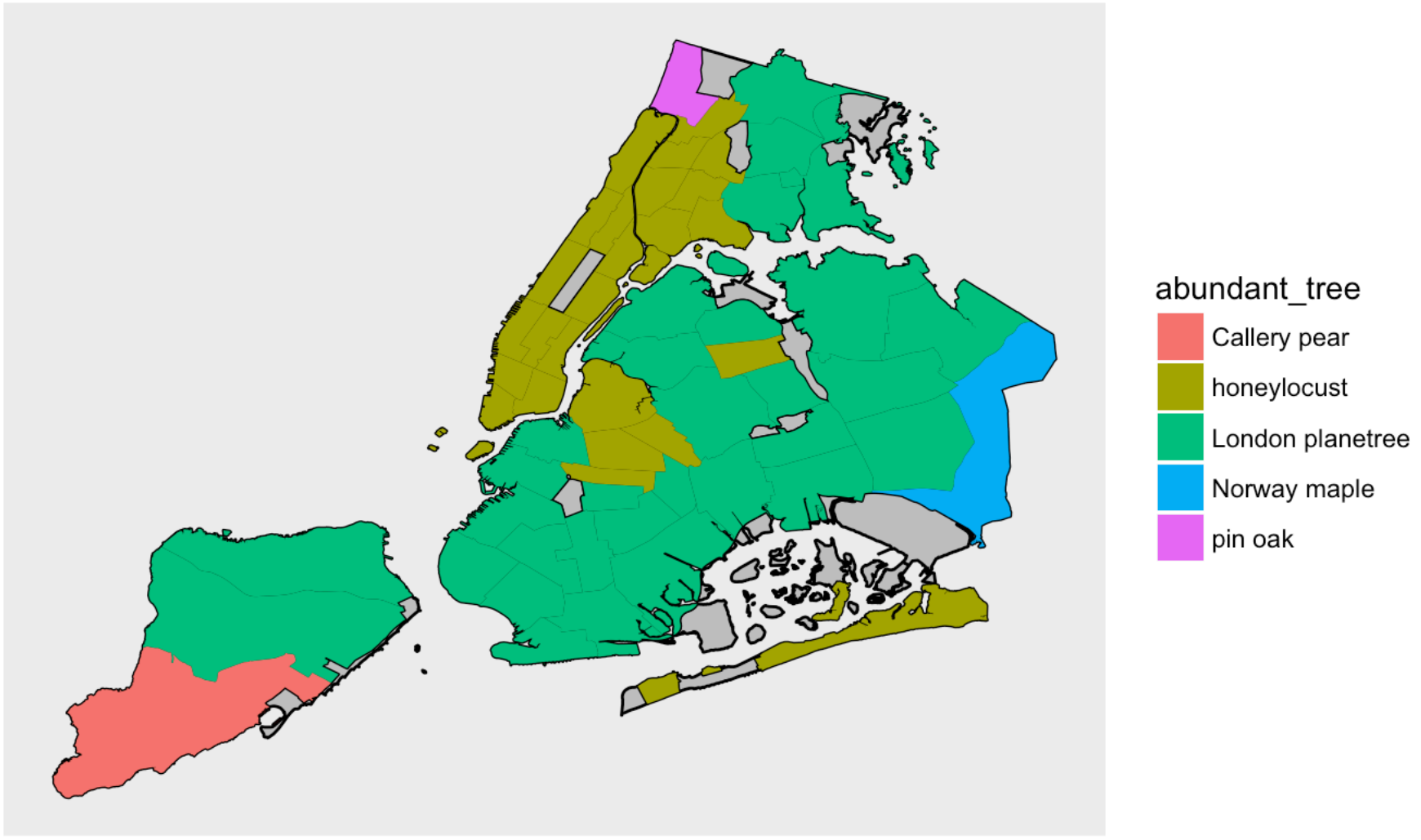


This map illustrates street tree diversity as the number of species in each district. From this visualization, it is apparent that Manhattan and Bronx districts contain fewer species of street tree on average. However, given that the community districts vary significantly in area, it might also be useful to plot this information per unit area by dividing num_species by each shape’s area.



This plot tells a different story - Manhattan and West Bronx are among the most diverse areas, judging by species per area. This may correspond to the earlier finding that Manhattan had the most trees per area while Staten Island had the least - more plantings is equated with greater diversity.

Moving on from diversity, are there any other interesting patterns at the species level? One idea might be to plot a map with the most common tree species in each district.



This plot displays the most abundant tree species in each district. The five species represented are the five most common overall, and the city-wide counts are as follows:

```
##
## London planetree      honeylocust     Callery pear          pin oak
##            87014            64263            58931            53185
##     Norway maple
##            34189
```

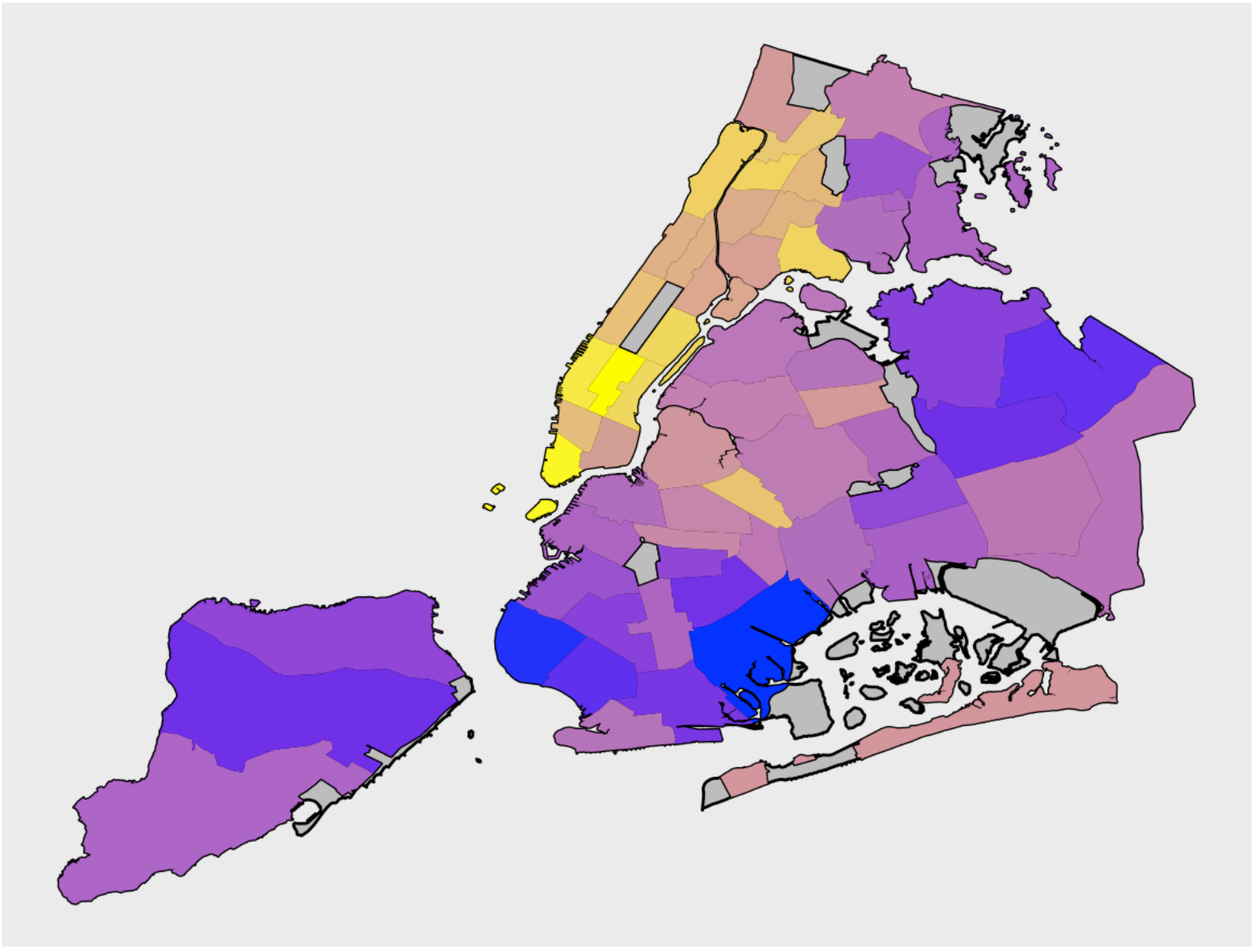
Given these numbers, the species distributions are quite interesting. Manhattan, West Bronx, and North Brooklyn are dominated by honey locusts, while the surrounding areas are predominantly planted with London planetrees. This seems to mirror the average trunk size plot created earlier. Is the planting distribution related to tree size?

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    6.00   10.00   10.21   13.00  169.00   31615
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   16.00   22.00   21.56   27.00  318.00   31615
```

Honey locust trees are about half as large (judging by diameter) as london planetrees, on average. This could mean that honey locusts are planted more often in the aforementioned congested parts of the city, while planetrees are better suited to more open neighborhoods.

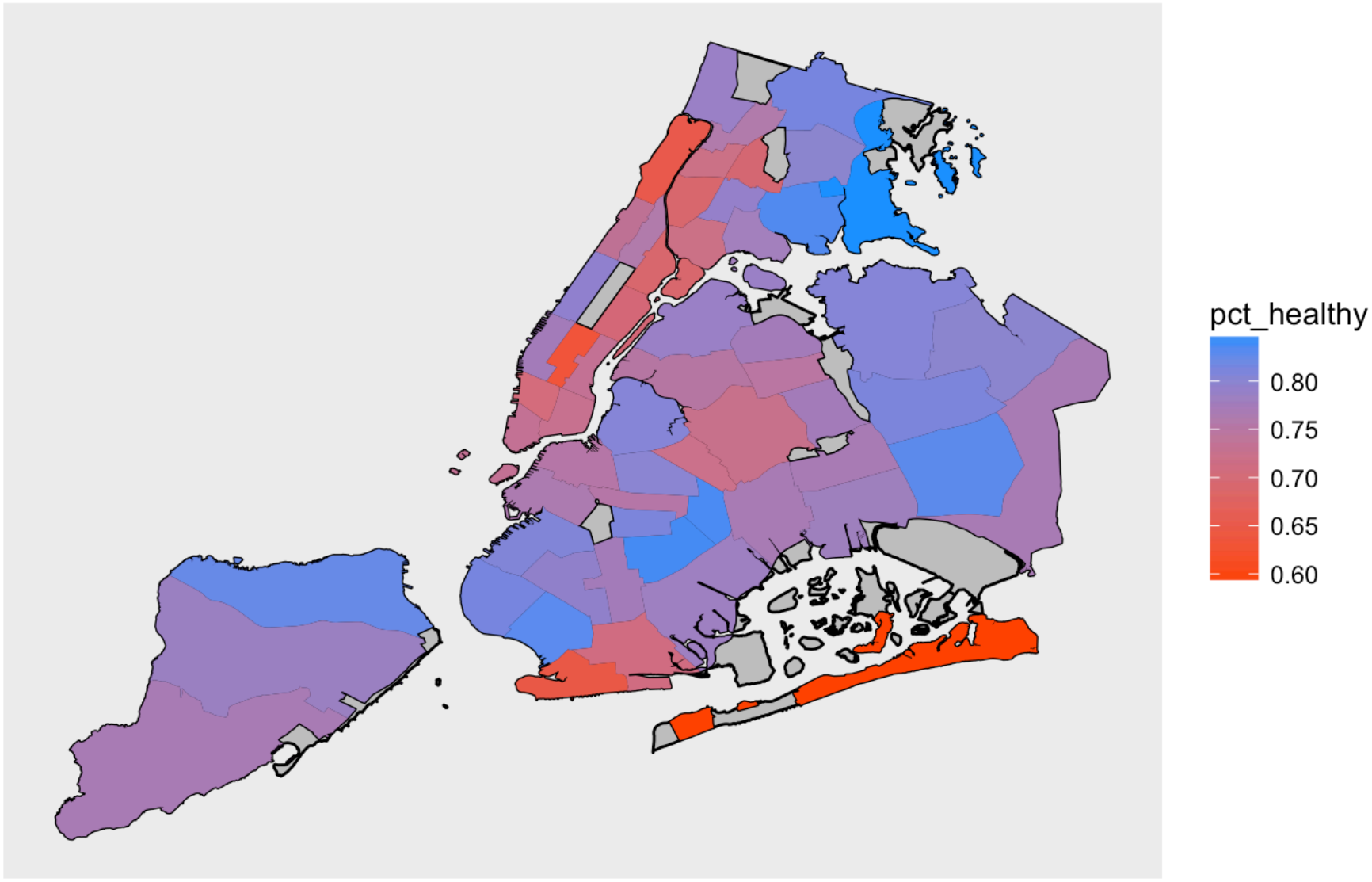
To better compare the distributions of the two species, we can make a map plot which only considers planetrees and honey locusts.



This plot shows the proportion of planetrees relative to honey locusts in each district. In much Manhattan, there are at least 4 honey locusts to every planetree (judging by the proportion of 0.2 on the legend). The ratio is closer in much of the rest of the city, with only a few districts colored blue or deep purple (indicating significant prevalence of planetrees).

To verify the assumptions I've been making about congestion in Manhattan and West Bronx, it might be a good idea to try to factor in some population data after I've had enough fun with these exploratory map plots.

Lastly, it would be interesting to revisit tree health in the context of location within the city.



This plot illustrates the proportion of healthy (health = ‘Good’) trees in each district. Much of Manhattan and West Bronx contains disproportionately unhealthy trees. Most notably, the bottom-right district - Rockaway, Queens - has a surprisingly low ‘health rating’. Only about 60% of trees were judged as being in good health. Why might this be the case? Bringing some more variables into the equation may help answer this question.

Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

Many variables exhibited interesting trends when related to location. Volunteer activity was much higher in the central parts of the city. Variables such as species diversity, tree size, and tree health seemed to have similar distributions with respect to location.

Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

The Rockaway peninsula in Queens had a very large percentage of unhealthy trees. This would be an interesting phenomena to explore further.

What was the strongest relationship you found?

The relationship between location and volunteer activity was quite profound, varying from less than 20% to over 80% depending on the community district.

Multivariate Plots Section

The bivariate exploration raised two location-related questions that will be further explored in this section:

1. Why do inner-city districts in Manhattan, West Bronx, etc. exhibit different species distributions and tree health?

2. Why are trees in Rockaway, Queens faring so poorly?

Starting with the first question, our intuition was that street trees fared worse in more congested districts. To justify this intuition, we can use another dataset provided by the city's OpenData initiative that includes populations for each of the 59 community districts, recorded once every decade (most recently in 2010). This data can then be joined to the communities shapefile and overlaid on the tree health plot using geom_point.

```
## 'data.frame':    59 obs. of  8 variables:
##  $ Borough         : Factor w/ 5 levels "Bronx","Brooklyn",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ CD.Number       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ CD.Name         : Factor w/ 59 levels "Astoria, Long Island City",..: 34 29 36
28 53 21 5 43 45 51 ...
##  $ X1970.Population: int  138557 99493 150636 144207 121807 114137 113764 103543 1
66442 84948 ...
##  $ X1980.Population: int  78441 34399 53635 114312 107995 65016 116827 98275 16762
7 106516 ...
##  $ X1990.Population: int  77214 39443 57162 119962 118435 68061 128588 97030 15597
0 108093 ...
##  $ X2000.Population: int  82159 46824 68574 139563 128313 75688 141411 101332 1678
59 115948 ...
##  $ X2010.Population: int  91497 52246 79762 146441 128200 83268 139286 101731 1722
98 120392 ...
```

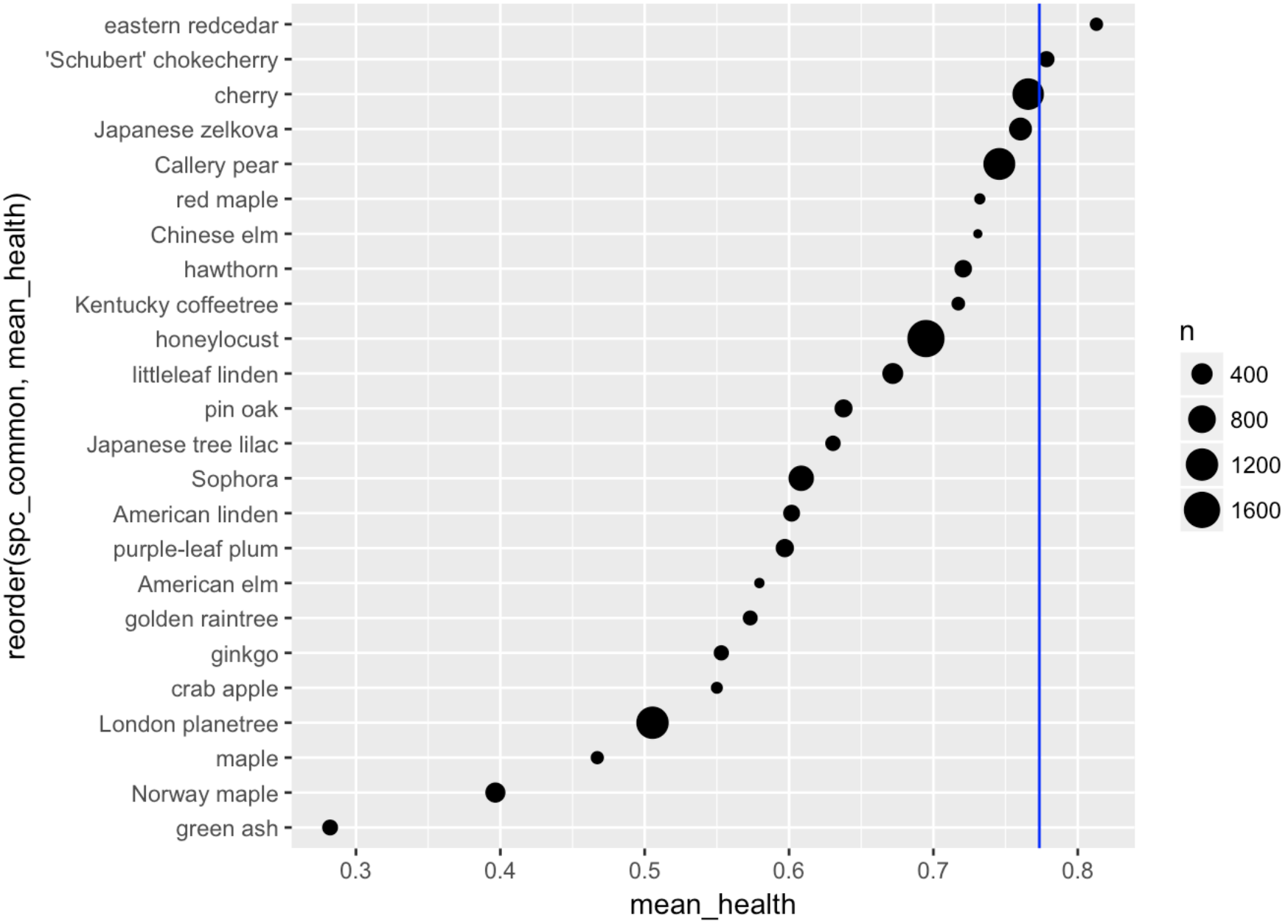
Judging by this plot, it seems that the theory of population density affecting tree health holds true in Manhattan and West Bronx. In those districts, the population densities are indeed very high compared to much of the rest of the city.

However, the relationship appears to be be weaker in Brooklyn and Queens - there are subtle trends in health that do not seem to directly related to population. The most striking example is the aforementioned Rockaway Peninsula - population density is quite low, yet the street tree health is worse on average than in any other district.

We can further investigate this district by subsetting the data and comparing tree health to more variables. Maybe some species tend to fare worse than others. A species-specific blight or insect infestation of an abundant species could bring down the average significantly.

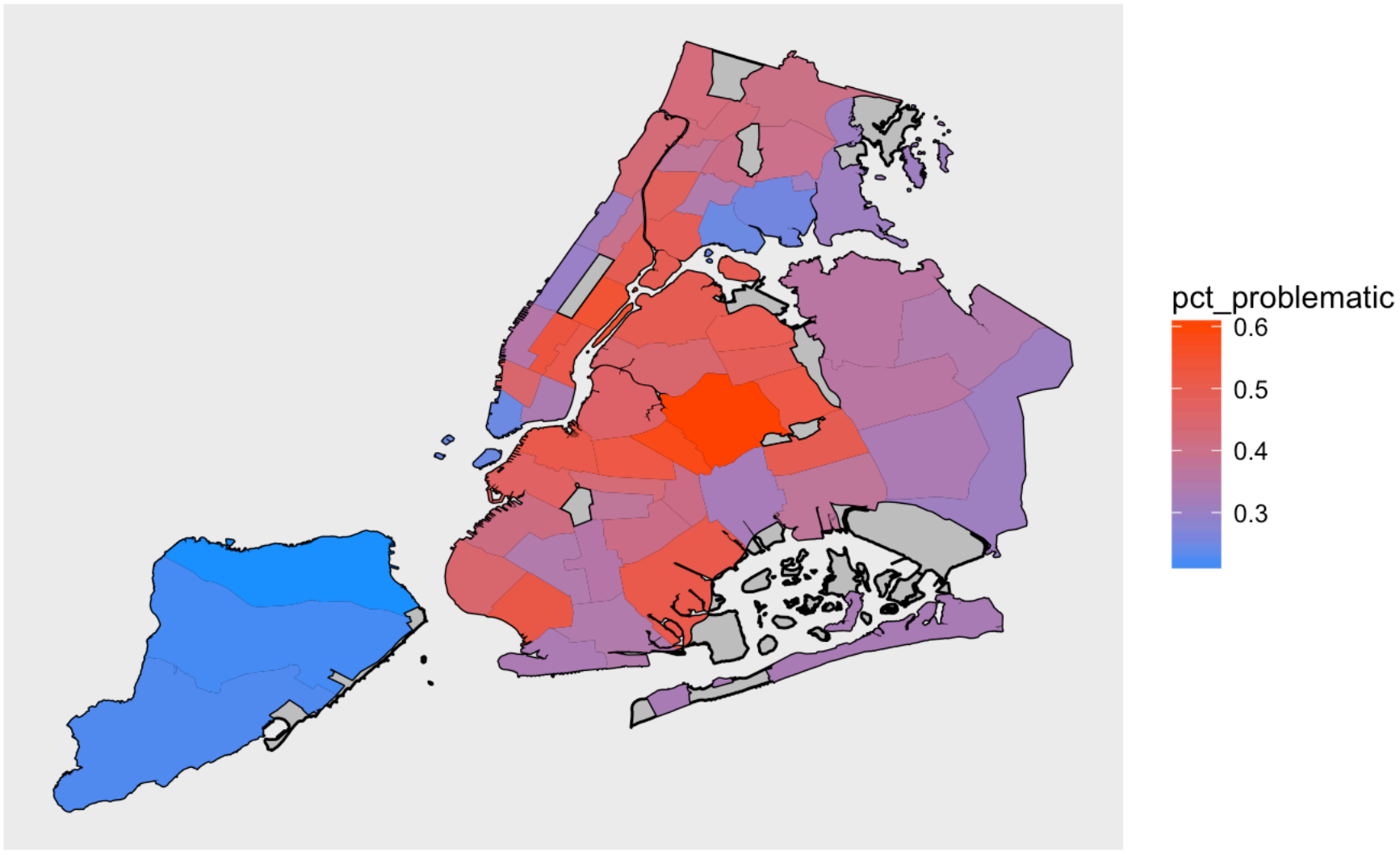
```
##    Mode   FALSE    TRUE    NA's
## logical      24     109        0
```

Rockaway is actually one of the most diverse districts, containing 109 of the total 133 species of NYC street tree. Let’s plot the average health for the most common species (over 100 individuals in the district).



This plot illustrates the average health of the most common species, with the dot size corresponding to tree count and the blue line representing the citywide average. (Again, health is measured as the number of live trees rated as being in ‘Good’ health divided by the total number of live trees.) Clearly, there is quite a bit of variation in health by species, but only two of the 24 species with counts over 100 exhibit health proportions greater than the citywide average. Notably, the two most common species in the city both fare poorly in Rockaway: about 50% of London planetrees and below 70% of honey locusts are in good health.

This plot indicates that many species of street tree suffer in this particular district, rather than one or two abundant species bringing down the average as theorized. An alternate theory is that the problems recorded for each tree, such as stones in planting beds and lighting affixed to the tree, occur in higher proportions in Rockaway. This can be investigated using another simple map plot.



Actually, Rockaway has a percentage of problematic trees on the lower end of the scale, so this theory doesn't hold up either.

The best remaining theory seems to be that Rockaway is simply an inhospitable environment for street trees. Perhaps storms are more likely to adversely affect the exposed trees in this beach-lined district. Another possibility is that young street trees are not as well cared for as described on the Parks Department webpage (https://www.nycgovparks.org/trees/street-tree-planting) during their first couple of years, a responsibility belonging to the street tree contractors.

Multivariate Analysis

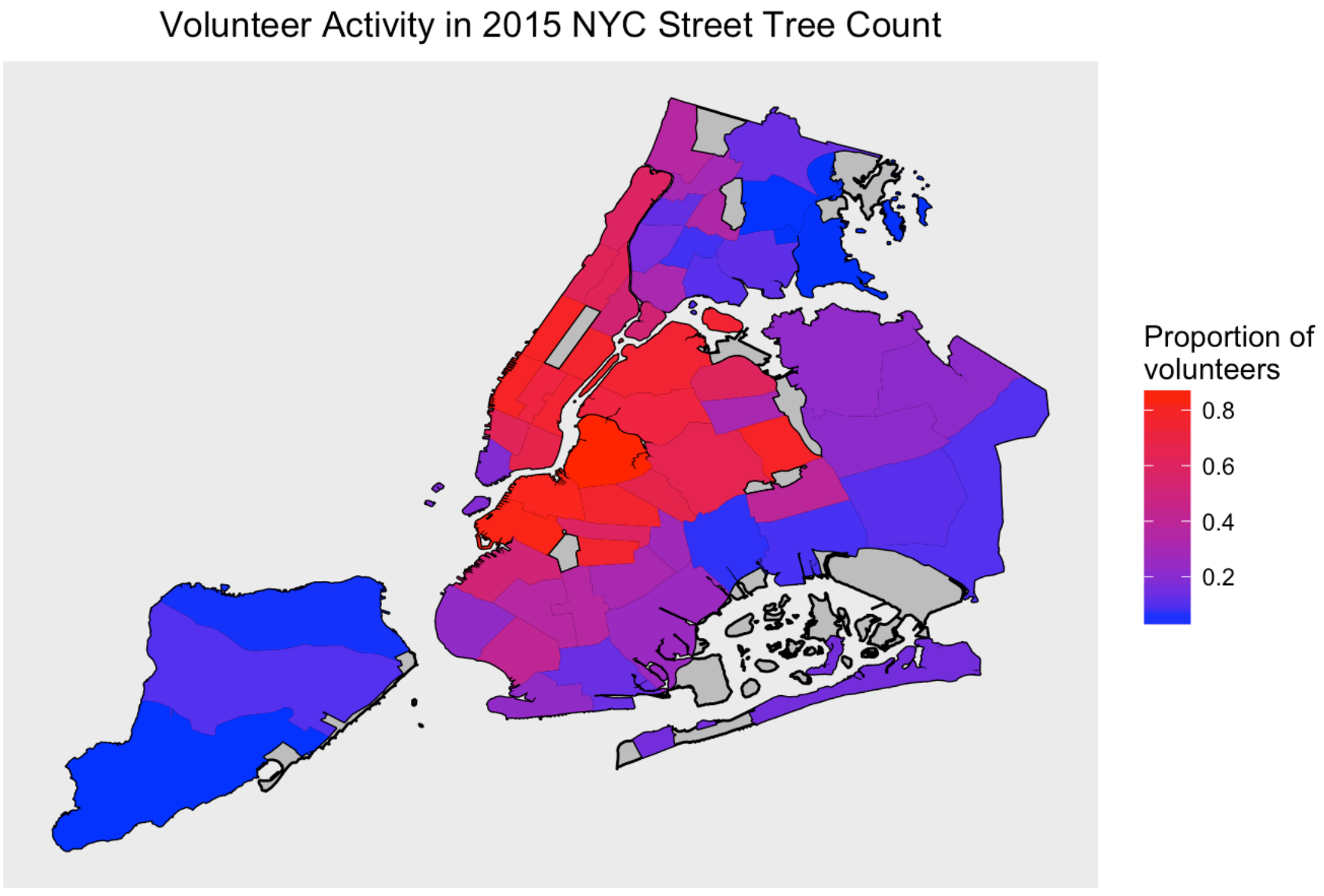
Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

The addition of the population data seemed to support the relationship between population density and tree health for Manhattan and Bronx, but the relationship was not as strong in other boroughs, the most notable outlier being the Rockaway Peninsula at the south of Queens. In Rockaway, poor tree health was found to be

representative of many of the common species, leading to theories outside the scope of this data.

Final Plots and Summary

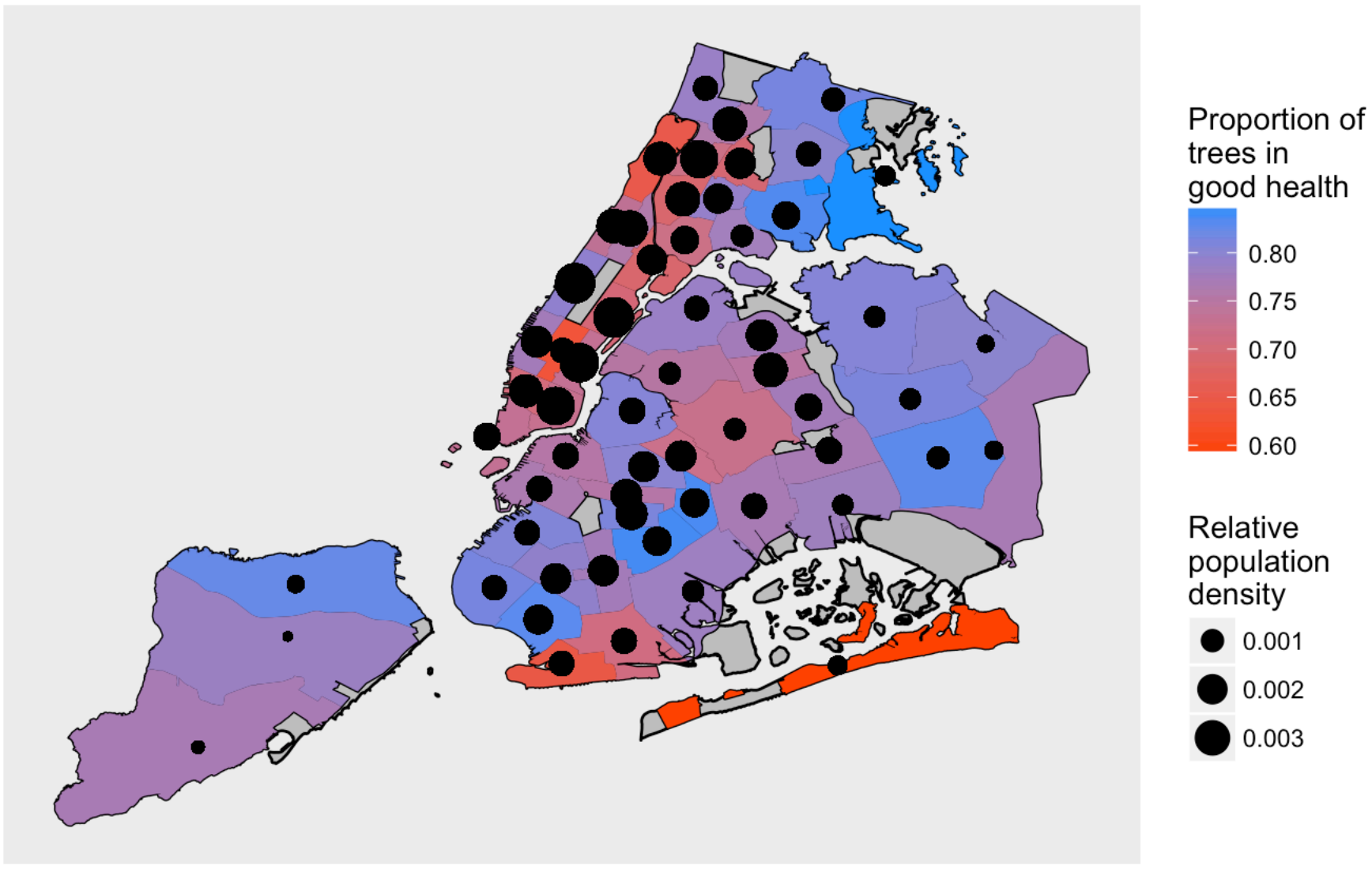
Plot One



Description One

This plot illustrates the proportion of trees counted by volunteers in each district during the 2015 street tree census. Volunteer activity varied significantly: central districts were catalogued almost entirely by volunteers, while Staten Island and outer Queens and Bronx were largely covered by staff. This plot could be used to determine areas in which volunteer engagement can be improved.

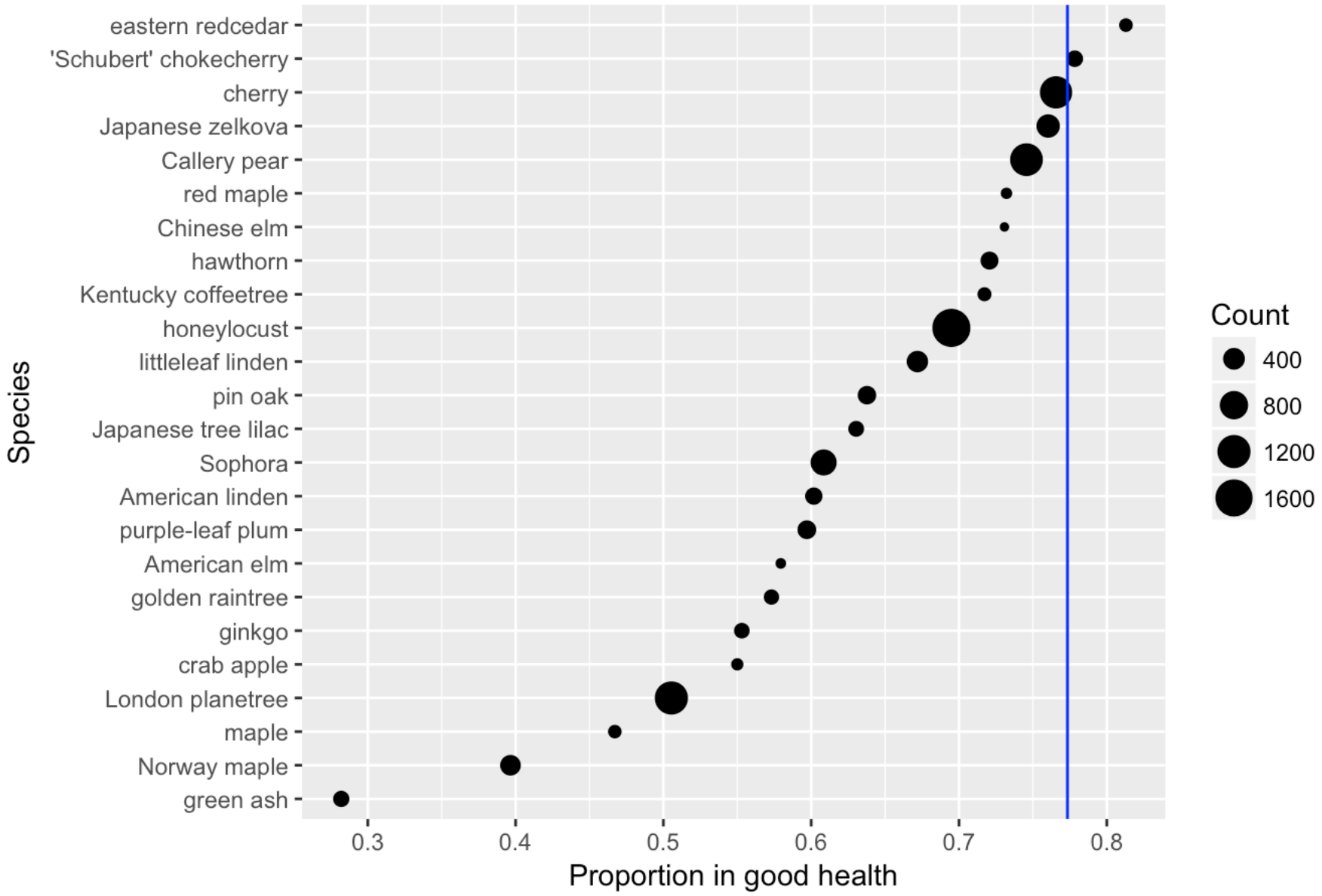
Plot Two



Description Two

This plot shows the proportion of trees in good health in each district, overlaid with dots whose size corresponds to the population density of each district. The relatively poor tree health in Manhattan and West Bronx seems to relate to the very high population densities of those neighborhoods. However, it appears to be possible to sustain healthy street tree populations even with high population density, as evidenced by the very healthy districts in central Brooklyn with moderately high population density. Lastly, the southeastern Rockaway district is a notable outlier.

Plot Three



Description Three

This plot illustrates the proportion of trees in good health for the most common species in the Rockaway district, with dot sizes corresponding to tree count and the citywide proportion plotted as the blue line. This plot was made to address the possibility that poor health in one or a few abundant species brought down the district's tree health rating. As it turns out, the overwhelming majority of the district's primary tree species fell below the citywide health rating, indicating that the problems with tree health in Rockaway are unlikely to be species-related.

Reflection

The most exciting parts of this investigation were learning to use map plots for compelling visualizations and leveraging data from two different datasets to make a stronger argument in the multivariate analysis. The function to generate the community district code for the population data was surprisingly difficult until I came across a Stack Overflow post (https://stackoverflow.com/questions/2641653/pass-a-data-frame-column-name-to-a-function) discussing how to pass dataframe column names to functions. I was interested in adding more socioeconomic data to investigate tree health in different districts, but wasn't able to find an appropriate dataset - future work might involve scraping additional information and adding it to map plots.