Data Import

Anthéa von Borowski, Mengjiao Li, Leonie Demsar

2024-12-09

Ressources

Libraries

```
#install.packages("sf")
#install.packages("dplyr")
#install.packages("ggplot2")
#install.packages("readr")
#install.packages("tidyverse")
#install.packages("Hmisc")
#install.packages("car")
#install.packages("here")

library(sf)
```

```
## Warning: Paket 'sf' wurde unter R Version 4.4.2 erstellt
```

```
library(dplyr)
```

```
## Warning: Paket 'dplyr' wurde unter R Version 4.4.2 erstellt
```

```
library(ggplot2)
```

```
## Warning: Paket 'ggplot2' wurde unter R Version 4.4.2 erstellt
```

```
library(readr)
library(tidyverse)
```

```
## Warning: Paket 'tidyverse' wurde unter R Version 4.4.2 erstellt

## Warning: Paket 'lubridate' wurde unter R Version 4.4.2 erstellt
```

```
library(Hmisc)
library(car)
library(here)
```

Raw Data

```
#ROHDATEN
#inc_zip<-read.csv(".\\Data\\Raw\\us_income_zipcode.csv")
#zcta<-st_read(".\\Data\\Raw\\ZCTA\\tl_2016_us_zcta510.shp")
#demog<-read.csv(".\\Data\\Raw\\ethn_zip.csv")
#data_nyc<-read.csv(".\\Data\\Raw\\baumnyc.csv")

# ALTERNATIVE MIT HERE()

inc_zip<-read.csv(here("Data", "Raw", "us_income_zipcode.csv"))
zcta<-st_read(here("Data", "Raw", "ZCTA", "tl_2016_us_zcta510.shp"))
```

```
## Reading layer 'tl_2016_us_zcta510' from data source
##   'C:\Users\anthe\OneDrive\Documents\Uni Stuttgart\Master\Data Science für Sozialwissenschaftler\Baum
##   using driver 'ESRI Shapefile'
## Simple feature collection with 33144 features and 9 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -176.6847 ymin: -14.37378 xmax: 145.8305 ymax: 71.34132
## Geodetic CRS:  NAD83
```

```
data_nyc<-read.csv(here("Data", "Raw", "baumnyc.csv"))
demog<-read.csv(here("Data", "Raw", "ethn_zip.csv"))
```

Mapping of locations to ZCTA zip-codes

```
points_ny <- st_as_sf(data_nyc, coords = c("longitude_coordinate", "latitude_coordinate"), crs = 4326) 

zcta_ny<-zcta[zcta$ZCTA5CE10 %in% 6390:14905,] #ignore all ZCTAs which are not NY
zcta_ny <- st_transform(zcta_ny, st_crs(points_ny)) #harmonize CRS values
zcta_ny$surface<-st_area(zcta_ny)

data <- st_join(st_sf(geometry = points_ny), zcta_ny, join = st_within) #join datasets to match zcta to
required_columns<-c("common_name","scientific_name","city", "state", "geometry", "zipcode", "condition"
baumnyc<-data[, required_columns, drop = F]
baumnyc$park<-is.na(baumnyc$ZCTA5CE10)

count_tree<-table(baumnyc$ZCTA)
count_tree
```

```
##
## 10001 10002 10003 10004 10005 10006 10007 10009 10010 10011 10012 10013 10014
##   779  2245  2044   218   122    50   255  1905   965  2183  1072  1110  2654
## 10016 10017 10018 10019 10020 10021 10022 10023 10024 10025 10026 10027 10028
##  1876   917   500  1688    69  1916  1526  2212  3327  3745  1657  3047  1723
## 10029 10030 10031 10032 10033 10034 10035 10036 10037 10038 10039 10040 10065
##  2379  1348  2610  2140  1998  1644  2064   902   723   441   855  1435  1833
## 10069 10075 10103 10110 10111 10112 10115 10128 10152 10154 10162 10165 10173
##   111   967    24     7    25    29     7  2318     8    16    29     3     4
## 10199 10278 10280 10282 10301 10302 10303 10304 10305 10306 10307 10308 10309
##     1     2   405   353  5832  2547  3179  6000  6798 13069  5380  7202 12471
```

```
## 10310 10311 10312 10314 10451 10452 10453 10454 10455 10456 10457 10458 10459
##  3575    35 22463 16767  2422  3361  3073  1708  2073  3914  3758  3319  3042
## 10460 10461 10462 10463 10464 10465 10466 10467 10468 10469 10470 10471 10472
##  3361  5611  4184  3926  1056  5149  5024  4185  2806  6917  1554  1795  3405
## 10473 10474 10475 10550 10704 10705 10803 11001 11004 11005 11020 11021 11040
##  4369  2674  1891     6     8     4     5  1467  4423    39    17    12   797
## 11101 11102 11103 11104 11105 11106 11109 11201 11203 11204 11205 11206 11207
##  3357  1920  2409  1625  3774  1955   189  4520  4993  4797  2530  4002  8622
## 11208 11209 11210 11211 11212 11213 11214 11215 11216 11217 11218 11219 11220
##  8450  6230  5186  6109  4235  3763  4336  5912  3499  3160  5044  4361  4897
## 11221 11222 11223 11224 11225 11226 11228 11229 11230 11231 11232 11233 11234
##  5035  3674  5919  1757  2971  3668  3629  6156  7519  3670  1857  4711 11355
## 11235 11236 11237 11238 11239 11354 11355 11356 11357 11358 11360 11361 11362
##  5403  6916  2962  4054   804  5698  5249  3114  9515  6874  2487  6250  4526
## 11363 11364 11365 11366 11367 11368 11369 11370 11371 11372 11373 11374 11375
##  2811  6977  7486  3320  5217  4469  3288  3061   150  3363  4190  3186  6957
## 11377 11378 11379 11385 11411 11412 11413 11414 11415 11416 11417 11418 11419
##  5539  3998  4886 10723  3307  4714  7443  4649  1685  1758  3691  3548  2812
## 11420 11421 11422 11423 11424 11426 11427 11428 11429 11430 11432 11433 11434
##  5484  3008  6304  3439    44  4643  4591  2961  2859    47  6869  3356  8447
## 11435 11436 11451 11580 11581 11691 11692 11693 11694 11697
##  4638  2343    66    13     1  5688  2087   784  3576    40
```

Processing Income Data

```
inc_zip<-inc_zip[inc_zip$Year ==2015,]
inc_zip$ZCTA<-as.numeric(substring(inc_zip$Geographic.Area.Name, 6))

inc_zip_ny<-inc_zip[inc_zip$ZCTA %in% 6390:14905,]
baumnyc$mean_income <- inc_zip$Households.Median.Income..Dollars.[match(baumnyc$ZCTA5CE10, inc_zip$ZCTA
```

Processing Ethnicities Data

```
#categories: white non hispanic/latino, black non hispanic/latino, asian non hispanic/latino, hispanic/

demog$ZCTA<-as.numeric(substring(demog$NAME, 7))
```

```
## Warning: NAs durch Umwandlung erzeugt
```

```
demog_ny<-demog[-1,]
demog_ny<-demog[demog$ZCTA %in% 6390:14905,]

baumnyc$pop<-as.numeric(demog_ny$B03002_001E[match(baumnyc$ZCTA5CE10, demog_ny$ZCTA)])
baumnyc$eth_white_nonhisp<-as.numeric(demog_ny$B03002_003E[match(baumnyc$ZCTA5CE10, demog_ny$ZCTA)])/bau
baumnyc$eth_afroam_nonhisp<-as.numeric(demog_ny$B03002_004E[match(baumnyc$ZCTA5CE10, demog_ny$ZCTA)])/ba
baumnyc$eth_asian_nonhisp<-as.numeric(demog_ny$B03002_006E[match(baumnyc$ZCTA5CE10, demog_ny$ZCTA)])/bau
baumnyc$eth_hisp<-as.numeric(demog_ny$B03002_012E[match(baumnyc$ZCTA5CE10, demog_ny$ZCTA)])/baumnyc$pop
baumnyc$eth_other<-1-(baumnyc$eth_white_nonhisp+baumnyc$eth_afroam_nonhisp+baumnyc$eth_asian_nonhisp+bau
baumnyc$count_tree<-count_tree[baumnyc$ZCTA5CE10]
```

```
baumnyc$count_tree[is.na(baumnyc$count_tree)]<-0
baumnyc$zcta_surface<-as.numeric(zcta_ny$surface[match(baumnyc$ZCTA5CE10, zcta_ny$ZCTA5CE10)])/1000000
baumnyc$pop_density<-baumnyc$pop/baumnyc$zcta_surface
baumnyc$tree_density<-baumnyc$count_tree/baumnyc$zcta_surface

coords<-st_coordinates
```

Recoding and grouping of data

```
baumnyc$native<-car::recode(baumnyc$native, "'naturally_occurring'= 0; 'introduced'=1;'no_info'=NA")
baumnyc$condition<-car::recode(baumnyc$condition, "'poor'=-1; 'fair' = 0; 'good' = 1")


baumnyc_gp<-data.frame(ZCTA = unique(baumnyc$ZCTA5CE10))

baumnyc_gp<- baumnyc %>%
  group_by(ZCTA = ZCTA5CE10) %>%
  dplyr::summarize(mean_native = mean(native, na.rm=T),
                   mean_condition = mean(condition, na.rm = T),
                   mean_income = mean(mean_income))

baumnyc_gp$pop<-as.numeric(demog_ny$B03002_001E[match(baumnyc_gp$ZCTA, demog_ny$ZCTA)])
baumnyc_gp$eth_white_nonhisp<-as.numeric(demog_ny$B03002_003E[match(baumnyc_gp$ZCTA, demog_ny$ZCTA)])/ba
baumnyc_gp$eth_afroam_nonhisp<-as.numeric(demog_ny$B03002_004E[match(baumnyc_gp$ZCTA, demog_ny$ZCTA)])/
baumnyc_gp$eth_asian_nonhisp<-as.numeric(demog_ny$B03002_006E[match(baumnyc_gp$ZCTA, demog_ny$ZCTA)])/ba
baumnyc_gp$eth_hisp<-as.numeric(demog_ny$B03002_012E[match(baumnyc_gp$ZCTA, demog_ny$ZCTA)])/baumnyc_gp$
baumnyc_gp$eth_other<-1-(baumnyc_gp$eth_white_nonhisp+baumnyc_gp$eth_afroam_nonhisp+baumnyc_gp$eth_asian
baumnyc_gp$count_tree<-count_tree[as.character(baumnyc_gp$ZCTA)]
baumnyc_gp$count_tree[is.na(baumnyc_gp$count_tree)]<-0
```