**SITO - DESCRIZIONI INTERESSANTI**

<https://docs.safegraph.com/v4.0/docs/neighborhood-patterns-2020>

(letto qua <https://www.safegraph.com/blog/announcing-neighborhood-patterns-safegraphs-newest-dataset> ha discorsi motivazionali carini tipo:

“using anonymized and aggregated location data to understand how devices move within *and* around CBGs can help businesses, researchers, and local governments better assess existing foot traffic trends and predict future consumer movement patterns”

e poi descrive cose interessanti per capire scopi potenzialità, un po’ forse già detti da Izzo)

[New York State Community Health Indicator Reports (CHIRS) (ny.gov)](https://health.ny.gov/statistics/chac/indicators/)

sito che dà info generali sulla popolazione di ogni county.

Da tenere a mente per la **pulizia del dataset**:

1. device\_home\_areas: Only cbgs with at least 2 devices are shown and cbgs with less than 5 devices are reported as 4
2. device\_daytime\_areas: Only cbgs with at least 2 devices are shown and cbgs with less than 5 devices are reported as 4
3. distance\_from\_home: If we have not identified a home area for at least 5 devices, distance\_from\_home will be blank
   1. dopo aver smattato per un’po per vedere quali sono le righe vuote mi sembra che siano tutte piene…
   2. ma il senso di questa feature? sarebbe quanto dista mediamente (o meglio come mediana) la CBG della riga da tutte le case entro 100m dei devices che vi sono andati nel mese?
4. distance\_from\_primary\_daytime\_location: If we have not identified a device\_daytime\_area for at least 5 devices, distance\_from\_primary\_daytime\_location will be blank
5. median\_dwell: Median dwell time in minutes. Note that we are only including stops that have a dwell of at least 1 minute
6. top\_same\_day\_brand: The value shown for each brand is a percentage representing: devices going to both the brand and the area / total devices stopping in the area.
7. census metadata e census block?
8. warning del coercing NA : con un ciclo for (riga 184 commentata del converter) abbiamo trovato che la top\_same\_day\_brand che risulta avere dei NA è la CBG “360050379004”, ovvero riga 14398 data set originale, 14384 del data set con righe tolte. il problema è che nel dataset originale non ha problemi quella variabile

**OBIETTIVI**:

1. avere tutte queste informazioni può essere utile per fare un’analisi del quartiere che quindi può essere poi considerata importante per definire il prezzo delle case, negozi,... Si potrebbe dire qualcosa di utile anche dato il covid no? ricordo che noi stiamo guardando giugno 2020 quindi magari i dati possono essere un po’ falsati.  
   <https://www.google.com/search?client=firefox-b-d&sxsrf=ALeKk01x2iM29xkJAHx5Wsef5CgUNyhM_w:1615024544976&q=june+2020+covid+cases+new+york&spell=1&sa=X&ved=2ahUKEwjhzLmis5vvAhUNs6QKHSVPCWQQBSgAegQIDRA1&biw=1366&bih=654>
2. chiedere di cosa ha bisogno Izzo:
   1. ha in mano un caso concreto (deve vedere dove aprire un bar ad esempio)
   2. non ha in mente niente e quindi proporre noi un’ipotesi:
      1. condividiamo un caso specifico di qualche azienda (tipo aprire un McDonald) -> analisi puntuale
      2. proporci come una società di consulenza che è in grado di modulare le analisi di identificazione delle migliori locations in base alle esigenze del cliente (esempio: mi focalizzo sulle colazioni -> guardo gli orari mattutini per aprire un bar vs mi focalizzo sulla sera per aprire un ristorante ecc) -> offrire una metodologia

From Neighborhood Patterns Dataset we want to analyze the geographic distribution of the CBGs, points of interests (shops, services) and time slots and predict permanence time of devices in CBG

**post call**

* regressione lineare, lasso cross validation cose sul primo obiettivo facile (+ chiedere la densità di popolazione dei metadata)
* è vero che i blocchi vicini con time stops simili formano una città? (non punto principale, secondo momento)
* dopo ridge regression: parliamo di classificazione e clustering -> sulle marche? / bisogna capire come sfruttare quei dati. suddividere i brand legati ad ambiti diversi. dividere i brand che abbiamo in classi, capire come interpretare qst dato. sai che in quella zona lì c’è o meno bisogno di una classe di brand(o di un brand specifico)
* su qst, occhio alla vicinanza tra i blocchi

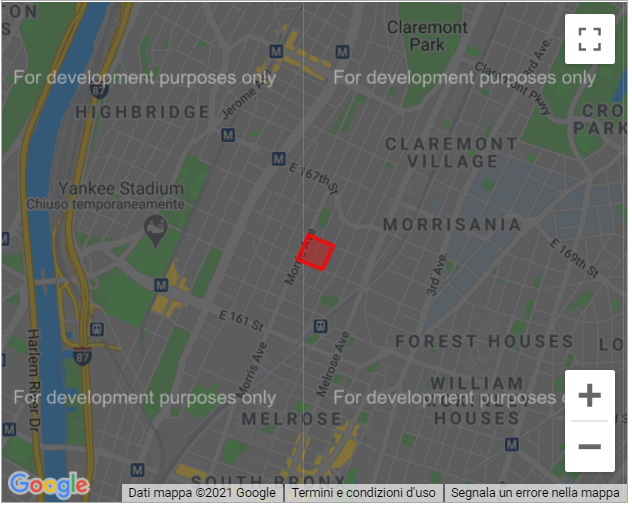
**PER PRESENTAZIONE**

* fare una mappatura grafica dei census blocks

**COME E’ FATTO IL DATASET?**

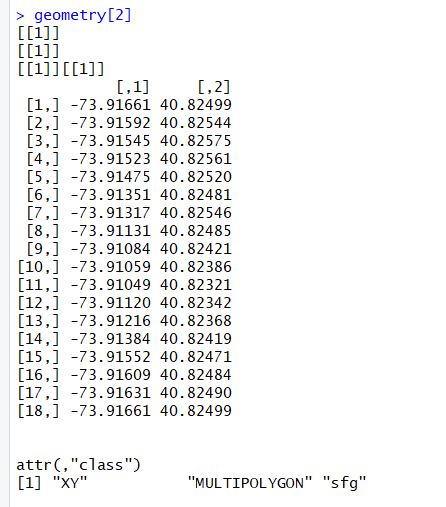
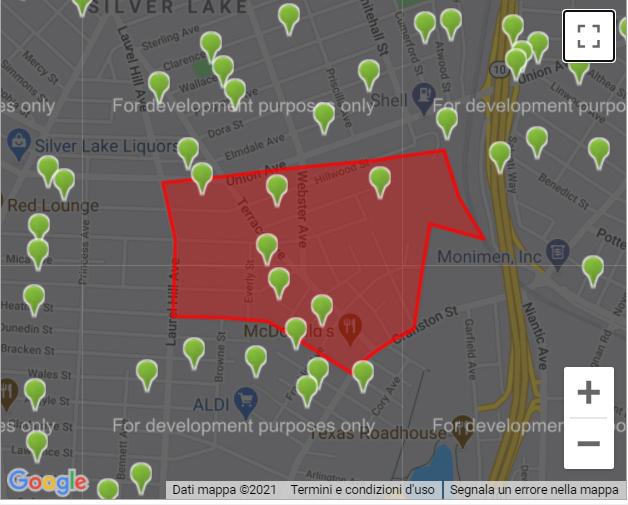
*census\_block\_ny*

* CensusBlockGroup = StateFIPS+CountyFIPS+TractCode+BlockGroup.
  + rappresenta l’isolato. messo su google trova l’identificazione esatta sulla mappa ([Census Block Group 017500-2 in Bronx County, New York (usa.com)](http://www.usa.com/NY0050175002.html))



dal sito possiamo trovare altre informazioni (etnie, isolato ricco o meno, ecc) utilizzabili quando faremo un’analisi specifica su alcuni isolati. utile per fare debugging per primi risultati (tipo un isolato malfamato non si ferma tanta gente)

definizione CBG sito \*A census block group is a geographic area defined by the United States Census Bureau and used for the census. On average, a census block group has around 1,500 residents. Census block groups, as well as census tracts, are more uniformly distributed in terms of the number of residents than cities or zip codes. Also, the census block group and the census tract demographic data are nearly 100% complete vs. less than 70% coverage of demographic data for cities and zip codes. Therefore census block groups and the census tracts are an excellent way to understand locations in a smaller scale, for example understanding the different areas of a large city. Census block groups are smaller than census tracts and can be further divided into census blocks for understanding locations at the block and community level.

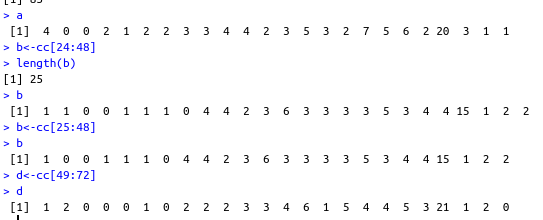
* State -> NY costante, non importante
* County -> quartiere
* class code -> area che racchiude più quartieri
* geometry -> da capire meglio. secondo me indica le coordinate dell’isolato:

*pattern\_ny*

* differenza tra raw\_stop\_counts e raw\_device\_counts:
  + il secondo è il numero delle persone (cell) che risiedono nell’area + le persone che entrano nell’area (complessivo nel mese) -> indice di densità di persone nell’area
  + il primo è il numero di fermate (>1 min). -> indicatore di movimento
  + esempio: vado a fare colazione nel bar sotto casa: il primo sono 2 fermate (dormo + mi siedo al bar), il secondo è 1.

chiediamogli se è effettivamente così o se 2 diversi stop sono considerati solo se si esce e poi si rientra nell’area (se spostarsi solo di qualche metro (quanto?) basta per far contare un altro stop allora uno che gira per 5 negozi diversi tutti attaccati viene considerato come 5 stop, nel caso di centri commerciali o simili potrebbe essere un problema)

* day\_counts strani dati (questi sono uguali per ogni riga, sono abbastanza inutili)
* capire conteggio stops\_by\_day vs stops\_by\_hour: mi aspetto che stops\_by\_day[1]= somma dei primi 24 valori di stops\_by\_hour. non torna



Non deve per forza tornare in quanto se uno si ferma per più di un’ora nello stesso punto (tipo a dormire) sarà contato come 1 stop in ogni ora ma conterà solo come 1 stop in stop\_by\_day

altra questione, lui conta dalle 16 alle 16 come 24h? (considerato che inizia a registrare dati alle 16, se mai volessimo interpretare i picchi)

* weekday\_device\_home\_areas -> utile per capire il numero di lavoratori che accedono a quell’area per lavorare + abitanti

weekday\_device\_home\_areas vs device\_home\_areas

* breakfast, lunch, afternoon, dinner sono le persone in un isolato in una parte del giorno
* work\_hours\_device\_home\_areas+work\_behavior\_device\_home\_areas -> dice quanto è attrattivo un isolato da un punto di vista lavorativo: lavoro residenziale vs passaggio diurno.
* da capire la differenza tra device\_daytime\_areas e work\_hours\_device\_home\_areas:

work\_hours\_device\_home\_areas -> home origin area + devices that from that home stopped in the CBG ros from 7.30 am to 5.30 pm

device\_daytime\_areas -> primary daytime origin areas + n devices from that origin area in hour 9am-5pm

o la differenza è semplicemente l’orario (secondo me sì)

oppure con “origin area” intende dire la CBG di origine generica, non necessariamente la casa. es primo valore prima riga dice che (works…home areas)CBG 60050073002 è la casa di 38 devices che sono stati dalle 7.30 alle 17.30 lì; mentre CBG 60050073002 è da dove provengono 50 devices prima di andare lì che consideriamo lavorare nella CBG della riga tra le 9-17. non so come sia possibile avere qst distinzione e cosa farcene …

|  |
| --- |
| * popularity\_by\_hour\_monday : sarebbe ora 1 ho 8 stops? così facile? Sì |

**NOTE UTILI PER R**

* credo che ogni volta che fai load di un dataset (tipo prima hai quello modificato da noi poi carichi quello originale), le varie tabelle (tipo pattern\_ny) si aggiornano in base al nuovo dataframe che hai caricato
* mi compare spess questo warning “ The following objects are masked from patterns\_ny (pos = 3):” nonostante prima abbia fatto detach … possible risposta da stackoverflow anche se mi convince poco ma R fa sbocco: “f you look at the down arrow in environment tab. The attached file can appear multiple times. You may need to highlight and run detach(filename) several times until all cases are gone then attach(newfilename) should have no output message.” (in effetti ho dovuto fare detach 4 volte finchè non ha detto che non esisteva patterns\_ny, anche se avevo fatto runnare solo una volta il codice e prima environment vuoto.

**colori**

<https://www.agnesevardanega.eu/wiki/r/grafici/gestione_colori>

<http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf>

<https://www.nceas.ucsb.edu/sites/default/files/2020-04/colorPaletteCheatsheet.pdf>

color <- hcl.colors(g1, palette = "Dark2") help palettes non ha funzionato molto

**sto usando questo**

[https://www.datanovia.com/en/blog/top-r-color-palettes-to-know-for-great-data-vi](https://www.datanovia.com/en/blog/top-r-color-palettes-to-know-for-great-data-visualization/#rcolorbrewer-palettes)

[sualization/#rcolorbrewer-palettes](https://www.datanovia.com/en/blog/top-r-color-palettes-to-know-for-great-data-visualization/#rcolorbrewer-palettes)

install.packages("RColorBrewer")

library(RColorBrewer)

|  |  |  |
| --- | --- | --- |
| CONTEA  [Albany](https://it.wikipedia.org/wiki/Contea_di_Albany_(New_York))  [Allegany](https://it.wikipedia.org/wiki/Contea_di_Allegany_(New_York))  [Bronx](https://it.wikipedia.org/wiki/Contea_del_Bronx)  [Broome](https://it.wikipedia.org/wiki/Contea_di_Broome) (204cbg)  [Cattaraugus](https://it.wikipedia.org/wiki/Contea_di_Cattaraugus)  [Cayuga](https://it.wikipedia.org/wiki/Contea_di_Cayuga)  [Chautauqua](https://it.wikipedia.org/wiki/Contea_di_Chautauqua_(New_York))  [Chemung](https://it.wikipedia.org/wiki/Contea_di_Chemung)  [Chenango](https://it.wikipedia.org/wiki/Contea_di_Chenango)  [Clinton](https://it.wikipedia.org/wiki/Contea_di_Clinton_(New_York))  [Columbia](https://it.wikipedia.org/wiki/Contea_di_Columbia_(New_York))  [Cortland](https://it.wikipedia.org/wiki/Contea_di_Cortland)  [Delaware](https://it.wikipedia.org/wiki/Contea_di_Delaware_(New_York))  [Dutchess](https://it.wikipedia.org/wiki/Contea_di_Dutchess)  [Erie](https://it.wikipedia.org/wiki/Contea_di_Erie_(New_York)) (città buffalo) 779cbg  [Essex](https://it.wikipedia.org/wiki/Contea_di_Essex_(New_York))  [Franklin](https://it.wikipedia.org/wiki/Contea_di_Franklin_(New_York))  [Fulton](https://it.wikipedia.org/wiki/Contea_di_Fulton_(New_York))  [Genesee](https://it.wikipedia.org/wiki/Contea_di_Genesee_(New_York))  [Greene](https://it.wikipedia.org/wiki/Contea_di_Greene_(New_York))  [Hamilton](https://it.wikipedia.org/wiki/Contea_di_Hamilton_(New_York))  [Herkimer](https://it.wikipedia.org/wiki/Contea_di_Herkimer)  [Jefferson](https://it.wikipedia.org/wiki/Contea_di_Jefferson_(New_York))  [Kings (Brooklin)](https://it.wikipedia.org/wiki/Contea_di_Kings_(New_York)) 2079 cbg  [Lewis](https://it.wikipedia.org/wiki/Contea_di_Lewis_(New_York))  [Livingston](https://it.wikipedia.org/wiki/Contea_di_Livingston_(New_York))  [Madison](https://it.wikipedia.org/wiki/Contea_di_Madison_(New_York))  [Monroe](https://it.wikipedia.org/wiki/Contea_di_Monroe_(New_York)) 609 cbg  [Montgomery](https://it.wikipedia.org/wiki/Contea_di_Montgomery_(New_York))  [Nassau](https://it.wikipedia.org/wiki/Contea_di_Nassau_(New_York)) (sud NY) 1143 cbg  [NewYork(Manhattan)](https://it.wikipedia.org/wiki/Contea_di_New_York)1168  [Niagara](https://it.wikipedia.org/wiki/Contea_di_Niagara)  [Oneida](https://it.wikipedia.org/wiki/Contea_di_Oneida_(New_York))  [Onondaga](https://it.wikipedia.org/wiki/Contea_di_Onondaga) 380 cbg  [Ontario](https://it.wikipedia.org/wiki/Contea_di_Ontario_(New_York))  [Orange](https://it.wikipedia.org/wiki/Contea_di_Orange_(New_York))  [Orleans](https://it.wikipedia.org/wiki/Contea_di_Orleans_(New_York))  [Oswego](https://it.wikipedia.org/wiki/Contea_di_Oswego)  [Otsego](https://it.wikipedia.org/wiki/Contea_di_Otsego_(New_York))  [Putnam](https://it.wikipedia.org/wiki/Contea_di_Putnam_(New_York)) (am NY)  [Queens](https://it.wikipedia.org/wiki/Contea_di_Queens_(New_York)) 1745 cbg  [Rensselaer](https://it.wikipedia.org/wiki/Contea_di_Rensselaer)  [Richmond (Staten Island)](https://it.wikipedia.org/wiki/Staten_Island)  [Rockland](https://it.wikipedia.org/wiki/Contea_di_Rockland)  [Saint Lawrence](https://it.wikipedia.org/wiki/Contea_di_St._Lawrence)  [Saratoga](https://it.wikipedia.org/wiki/Contea_di_Saratoga)  [Schenectady](https://it.wikipedia.org/wiki/Contea_di_Schenectady)  [Schoharie](https://it.wikipedia.org/wiki/Contea_di_Schoharie)  [Schuyler](https://it.wikipedia.org/wiki/Contea_di_Schuyler_(New_York))  [Seneca](https://it.wikipedia.org/wiki/Contea_di_Seneca_(New_York))  [Steuben](https://it.wikipedia.org/wiki/Contea_di_Steuben_(New_York))  [Suffolk](https://it.wikipedia.org/wiki/Contea_di_Suffolk_(New_York)) 999 cbg  [Sullivan](https://it.wikipedia.org/wiki/Contea_di_Sullivan_(New_York))  [Tioga](https://it.wikipedia.org/wiki/Contea_di_Tioga_(New_York))  [Tompkins](https://it.wikipedia.org/wiki/Contea_di_Tompkins)  [Ulster](https://it.wikipedia.org/wiki/Contea_di_Ulster)  [Warren](https://it.wikipedia.org/wiki/Contea_di_Warren_(New_York))  [Washington](https://it.wikipedia.org/wiki/Contea_di_Washington_(New_York))  [Wayne](https://it.wikipedia.org/wiki/Contea_di_Wayne_(New_York))  [Westchester](https://it.wikipedia.org/wiki/Contea_di_Westchester) (amNy) 704  [Wyoming](https://it.wikipedia.org/wiki/Contea_di_Wyoming_(New_York))  [Yates](https://it.wikipedia.org/wiki/Contea_di_Yates)  BoloTown | ANOVA p-value  1 alto cazzo troppo  1  0.158  0.0335 \*  0.969  0.993  0.792  0.469  0.995  0.938  0.999  0.929  0.999  0.941  0.0122 \*  0.999  0.993  0.992  0.948  0.998  0.89  0.974  0.999  1.2e-08 \*\*\*  0.999  0.995  0.998  0.0646 . sni  0.999  4.64e-06 \*\*\*  0.968  0.694  1  0.0826 . sni  0.975  0.848  0.999  0.956  0.992  0.992  3.94e-05 \*\*\*  0.896  0.277  0.767  0.952  0.92  0.598  1  0.997  1  0.885  0.00139 \*\*  0.996  0.998  0.997  0.971  0.996  1  0.998  0.0792 . (sni)  1  1  fa cacca | bartlett test (for var)  p-value = 0.004083  p-value = 0.8705  p-value < 2.2e-16  p-value = 2.615e-07  p-value = 0.0462  p-value = 3.434e-10  p-value = 0.03807  p-value = 0.001964  p-value = 0.1273  p-value = 0.005345  p-value = 0.7601  p-value = 0.2039  p-value = 0.7553  p-value = 1.704e-05  p-value = 4.267e-09  p-value = 0.6084  p-value = 0.3551  p-value = 0.6451  p-value = 0.2145  p-value = 0.1169  p-value = 0.7068  p-value = 0.1973  p-value = 2.248e-09  p-value < 2.2e-16  p-value = 0.0147  p-value = 0.8425  p-value = 0.4721  p-value < 2.2e-16  p-value = 0.9035  p-value = 1.1e-05  p-value < 2.2e-16  p-value < 2.2e-16  p-value = 0.2147  p-value = 2.913e-08  p-value = 2.826e-06  p-value = 0.03989  p-value = 0.8002  p-value = 0.03347  p-value = 0.1114  p-value = 0.9559  p-value = 6.653e-09  p-value = 0.1985  p-value = 0.3808  p-value < 2.2e-16  p-value = 0.0004416  p-value = 3.994e-05  p-value = 0.2631  p-value = 0.7894  p-value = 0.7185  p-value = 0.9397  p-value = 0.06936  p-value = 0.0002852  p-value = 0.1108  p-value = 0.6595  p-value = 0.3403  p-value = 0.0509  p-value = 0.6256  p-value = 0.9832  p-value = 0.8493  p-value = 2.299e-05  p-value = 0.8572  p-value = 0.9618 |

CITTÀ PIÙ DENSE DI POPOLAZIONE

1. [New York](https://it.wikipedia.org/wiki/New_York), 8 175 133
2. [Buffalo](https://it.wikipedia.org/wiki/Buffalo), 261 310 (capoluogo contea erie)
3. [Rochester](https://it.wikipedia.org/wiki/Rochester_(New_York)), 210 565 (capoluogo contea monroe)
4. [Yonkers](https://it.wikipedia.org/wiki/Yonkers), 195 976 (città contea [Westchester](https://it.wikipedia.org/wiki/Contea_di_Westchester))
5. [Syracuse](https://it.wikipedia.org/wiki/Syracuse_(New_York)), 145 170 (capoluogo [Onondaga](https://it.wikipedia.org/wiki/Contea_di_Onondaga))
6. [Albany](https://it.wikipedia.org/wiki/Albany_(New_York)), 97 856 (capoluogo albany)
7. [New Rochelle](https://it.wikipedia.org/wiki/New_Rochelle), 77 062 (città [Westchester](https://it.wikipedia.org/wiki/Contea_di_Westchester))
8. [Mount Vernon](https://it.wikipedia.org/wiki/Mount_Vernon_(New_York)), 67 292 (città [Westchester](https://it.wikipedia.org/wiki/Contea_di_Westchester))
9. [Schenectady](https://it.wikipedia.org/wiki/Schenectady), 66 135 (città di [Schenectady](https://it.wikipedia.org/wiki/Contea_di_Schenectady))
10. [Utica](https://it.wikipedia.org/wiki/Utica_(New_York)), 62 235 (capoluogo oneida)

|  |  |  |
| --- | --- | --- |
|  |  |  |