slide 1

Hi everybody today we want to talk about cyber capital. If you expect to find data describing the movement of each person captured with a personal smartphone, (for instance where a person comes from and goes to, which bar or pub visits ecc, ….. ) don’t worry this dataset gives only local information upon geographic areas. so our privacy is safe for this time.

slide2

These geographic areas are called Census Block Groups and each one belongs to a county the state of NY.

CBG are smaller than cities and zip codes. they’re very useful because they give insights and movements about different areas to understand more precise locations of a large city.

Starting from a total of 15461 cbgs, we discarded 15 units while we were making the dataset more “R readable”.

slide 3

For each cbg we have 35 features. Some of them are in common to all, such as the start and end of the data measurement, the state and region those cbg belong to and count of days in june 2020.

[mese giugno si erano allentate le misure covid quindi può essere considerato un mese normale per quanto riguarda gli spostamenti.]

slide 6 -

Other more relevant features are related to the single cbg: for example median\_dwell could be considered as the characteristic time of stay of a person(a device) in the cbg.

[Median dwell time in minutes. Note that in all our dataset we are only including stops that have a dwell of at least 1 minute.]

slide 7

Other features give insights about the movements between cbgs. In general, they tell how many devices come from a specific cbg-home in the cbg of analysis and at which time of the day(circa). (**quanto vogliamo dire di qst? può essere detto meglio sicuro)**

**brand? mi sembra li balziamo, speriamo al massimo qualcuno chieda**

slide 8

[“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”]

Having such information our main aim is to understand how many people are attracted by a specific cbg in order to predict future consumer movement patterns.

To do so we want to analyse which are the “busiest” days, if cbgs could be divided between city and countryside according to a density threshold and so which cbgs are more attractive and when.

Moreover we plan to do a **linear regression to determine the time of stay**(median\_dwell) of a person in a cbg [ and trying to find according to which features it changes and try to interpret why. (elenco esempi di features ...)]

Upon this information and feature related to brands, we would like to help allocating brands **[questo bah …forse non lo direi]**

slide 9

We tried to give a first look to our dataset using anova, as a preliminary analysis. We weren’t able to verify the gaussianity assumptions because we have too many statistical units(so shapiro test was not usable), moreover homoscedasticity of variances is not verified. So we are aware of the fact that we need to improvethe modeling assumptions and settings.

slide 10

We did a one-way anova to analyze the variance of mean density of stops grouped with respect to the 7 days of the week or {week-end , work} days. Qualitatively it seems to show a sensible difference but since the pvalue is too high we are not able to refuse H0 hypothesis and the means od the mean density of stops are to be considered equal.

slide 11

We also did a two-ways anova grouping the density of stops in the cbgs with respect to the 62 counties of the NY state and the week days. It showed no interaction between days and counties.

slide 12

Showing a big difference in the means related to the county the cbg belongs to, we decided to focus our analysis on NY city: indeed all her counties have a higher density of population with respect to the others and it is of more “business” interest.

There are 5 counties .. nomi ...

slide 13

Also in this case anova assumption of gaussianity of groups and homoscedasticity of variances are not verified, but still we try to be driven in our approach to the dataset by those results. INdeed as expected from the graphs it is shown that there is no significant interaction between days and NY counties and still in the additive model is more relevant the information on the county.

slide 15

Actually we analyse two types of variables (not shown all): a density and a “pure” number of stops. We still need to understand if it could be better to consider a CBG as a single point or an area(so with spatial information also)

Moreover since we are dealing with a lot of features we should look for a way to “group” or select them. By similarity between some we could try PCA, for knowledge on the counties(cities or country) we could try clustering.

commenti a caso

- CHIEDI A LOMBA UNITÀ DI MISURA DENSITÀ(ANCHE RISPETTO ALLA DEF DI CBG STO INIZIANDO AD AVERE DUBBI ..) L’unità di misura sono i km^2

[(Our dataset is taken from SafeGraph site and Neighborhood Patterns contains footfall data aggregated by census block group. )

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.

Its information has been collected during the month of june 2020 and is related to the state of NY; for privacy reasons those data are related only to the cbg. A census block group is a ...(DEF)

(far rendere con semplicità l'idea di scala,dallo stato di NY alla contea, al cbg)

questa definizione troppo formale più scialla

\*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.]

“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”

* 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
* cluster: raggruppi i cbg come densità popolazione, periferia o no -> magari per trovare un threshold della densità per capire cosa è città e cosa no (di per sè inutile .. però vd mail izzo)