**Week 3 Thursday: Pre-class R exercise**

In Thursday’s class you’ll be exploring a dataset from the Urban Environment and Social Inclusion Index, which provides indicators for how well cities around the world perform on key environmental issues like air quality, climate change, tree cover, and public transit.

You can read more about the data in detail (<http://www.datadrivenlab.org/urban>) if you’re interested or explore definitions of the data in the ‘UESI\_indicator\_metadata.pdf’ document on Canvas.

Included in the spreadsheet we’ll explore on Tuesday are some general information about the various cities (population, population density, area, etc.) as well as values of air pollution, urban heat island, tree cover, and public transportation.

Higher values for indicators like air pollution (PM25\_mean; NO2\_mean) and Urban Heat Island (UHI\_mean) are bad; for others like (TREECAP\_mean; PUBTRANS\_mean; TRANSCOV\_mean) have higher values on the good end. Also included are normalized indicator scores (translating the raw values to a scale of 0 to 100, where 100 is `at target` or means that the city has the highest score possible). The same indicators are denoted with `\_UESI`.

In this document are a few questions about the data. Working through this preliminary exploration will help prepare you to answer some more questions during class. For class on Thursday, **please have ready-to-run R code that will answer these questions**. It doesn't have to be the most elegant or 'best' way to do things; it just needs to work. We’ll talk about alternatives in class, and introduce you to some nifty new things you can do in R.

We’ll get you started: (Note: the R-code in boxes below may or may not work properly if cut-and-pasted because of issues with formatting of special characters in Word; in any case, typing these commands yourself can help you get more comfortable with R syntax.

uesi <- read.csv("UESI2019\_with\_indicators.csv")

Explore the data as usual with

dim(uesi)

head(uesi)

str(uesi)

table(uesi$continent)

unique(uesi$country)

**Investigating missing values**

'Missing values (NA) are a challenge to handle properly in any language. Think of missing (NA) as "I don't know". Suppose a value for a particular city is missing. Then what is the answer to the question

"Is the value equal to 100?" The answer: we really don't know! So we sometimes need to tell R when it is okay to ignore missing values. It could be that some of the missing values are actually 100, but of course we don't know. We saw in Week 3 Monday’s class a little bit how NA or missing values, if we don’t tell R to remove them, it can return an error.

Let’s say you wanted to find the median income of cities in the UESI:

R returns ‘NA’ because there are missing values!

median(uesi$income\_mean)

Including the argument ‘na.rm=T’ fixes this problem:

median(uesi$income\_mean, na.rm=T)

But perhaps you want to explore how representative the median is, given the number of NA values. We can use the table() function as well as a new function is.na() to investigate:

table(is.na(uesi$income\_mean)

**Introducing logical and relational operators in R**

You’ve already been introduced to some logical operators (like &) in R. R can also understand a range of other logical and relational operators:

|  |  |
| --- | --- |
| Operator | Description |
| < | Less than |
| <= | Less than or equal to |
| != | Not equal to |
| ! | Logical not |
| | | Logical OR |
| & | Logical AND |

These can be helpful to help answer questions like:

* Total population of cities not in Asia.
* The number of cities with more than 100 neighborhoods
* Which cities have scores above 85 on both PUBTRANS.UESI and

TREECAP.UESI

See how many of the above questions you can answer in R using logical and relational operators, plus functions you’ve learned before.