Department of City and Regional Planning
CPLN 505 Planning By Numbers/ Spring 2019
Professor: Megan Ryerson
TA: Joshua Davidson

Assignment 2: Comparing Groups and Testing for Association and Correlation

(due February 22, 2019, 11:59 pm)

Philadelphia is served by a Combined Sewer System (CSS) that was originally constructed during the Industrial Revolution, over 100 years ago. In an effort to upgrade its infrastructure to meet 21st century environmental standards, the city requires significant investment in stormwater and wastewater infrastructure. The previous administration, headed by Mayor Michael Nutter, was successful in adopting one of the most aggressive green infrastructure plans in the country as a means to reduce loading to the system, while also providing economic and environmental benefit to the city. Mayor Kenney has brought you on as a consultant of the City of Philadelphia to evaluate how Philadelphia compares to other cities in water infrastructure needs. To do this, you need to characterize and compare the water infrastructure needs across the entire United States.

The Clean Watersheds Needs Survey (CWNS) is a publicly available dataset containing wastewater utilities' self-reported facility budgetary needs. Survey results have been released for 2004, 2008, and 2012. You are provided with cleaned data from the 2012 survey, which has been spatially joined with county-level data from the US Census.

For students who would like to map this data in GIS software, latitude and longitude data are provided. Export the provided .csv (file to .dbf format using the R package `foreign`. In ArcGIS, Add the .dbf file, then in the Layer tree, right click on the tabular dataset, and select 'Display XY Data'. When selecting the spatial reference, choose: 'Geographic Coordinate System NAD 1983 (2011)'**.** From here, you can do spatial joins to other datasets (you don't have to for this assignment, but this is often very useful).

Data Key:

Column Name	**Description**	**Source**
STCOU	State and county identifier	US Census County-level data
CWNS_NUMBER	Facility identifier	EPA 2012 CWNS
TOTAL_OFFICIAL_NEED	Total reported need for water infrastructure	EPA 2012 CWNS
PRES_RES_REC_COLLCTN	Present (2012) residential population receiving collection	EPA 2012 CWNS
PRES_RES_REC_TRMT	Present (2012) residential population receiving treatment	EPA 2012 CWNS
PRES_N_RES_REC_COLLCTN	Present (2012) non-residential population receiving collection	EPA 2012 CWNS

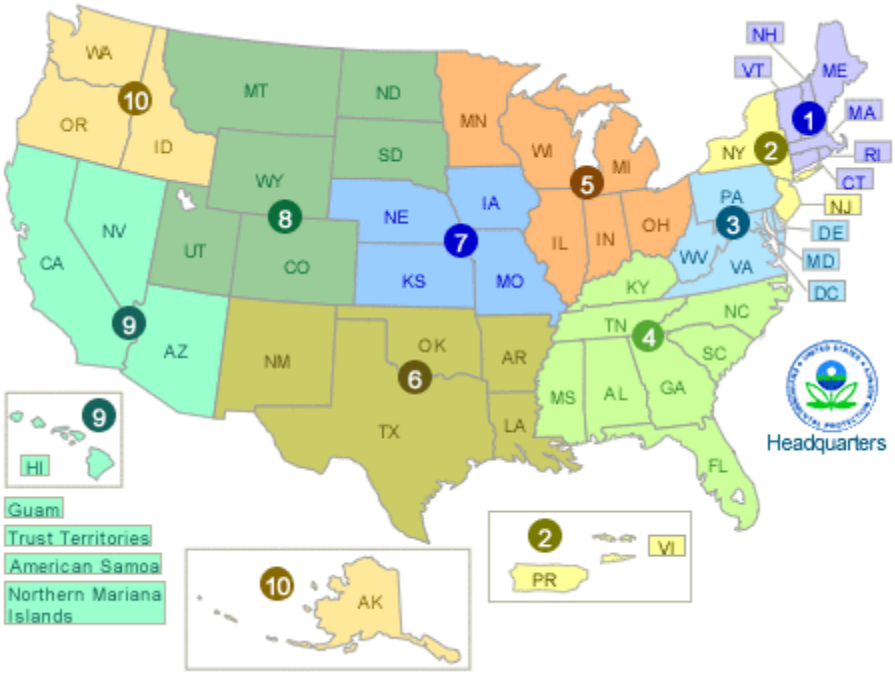
PRES_N_RES_REC_TRTM	Present (2012) non-residential population receiving treatment	EPA 2012 CWNS
PROJ_RES_REC_COLLCTN	Projected residential population receiving collection	EPA 2012 CWNS
PROJ_RES_YR	Projection year for residential population	EPA 2012 CWNS
PROJ_RES_REC_TRMT	Projected residential population receiving treatment	EPA 2012 CWNS
PROJ_N_RES_REC_COLLCTN	Projected non-residential population receiving collection	EPA 2012 CWNS
PROJ_N_RES_YR	Projection year for non-residential population	EPA 2012 CWNS
PROJ_N_RES_REC_TRMT	Projected non-residential population receiving treatment	EPA 2012 CWNS
EPA_REGION	EPA Region (1-10), see map below	EPA 2012 CWNS
STATE	State where facility is located	EPA 2012 CWNS
FACILITY_NAME	Name of facility	EPA 2012 CWNS
OWNER_TYPE	Public, Private, Federal	EPA 2012 CWNS
FACILITY_CITY	City where facility is located	EPA 2012 CWNS
LATITUDE	Latitude of facility	EPA 2012 CWNS
LONGITUDE	Longitude of facility	EPA 2012 CWNS
TMDL_INDICATOR	Y=contributing to a TMDL receiving water body	EPA 2012 CWNS
STATEFP	State code	US Census County-level data (via spatial join)
COUNTYFP	County code of location of facility	US Census County-level data (via spatial join)
COUNTYNAME	County name of location of facility	US Census County-level data (via spatial join)
ALAND	Land area (m^2) of county	US Census County-level data (via spatial join)
AWATER	Water area (m^2) of county	US Census County-level data (via spatial join)
MEDINC69	Median income in 1969 in county	US Census County-level data (via spatial join)
MEDINC79	Median income in 1979 in county	US Census County-level data (via spatial join)
MEDINC89	Median income in 1989 in county	US Census County-level data (via spatial join)
MEDINC99	Median income in 1999 in county	US Census County-level data (via spatial join)
MEDINC09	Median income in 2009 in county	US Census County-level data (via spatial join)
MEANINC09	Mean income in 1969 in county	US Census County-level data (via spatial join)
POP80	Total population in county in 1980	US Census County-level data (via spatial join)
POP90	Total population in county in 1990	US Census County-level data (via spatial join)
POP00	Total population in county in 2000	US Census County-level data (via spatial join)
POP10	Total population in county in 2010	US Census County-level data

PCTWHITE00	Percent of county population identifying as white in 2000	US Census County-level data (via spatial join)
PCTWHITE10	Percent of county population identifying as white in 2010	US Census County-level data (via spatial join)
CSS	1=facility associated with a combined sewer system (CSS)	EPA 2012 CWNS (via T. Lim recategorization)
MS4	1=facility associated with a municipal separate sewer system (MS4)	EPA 2012 CWNS (via T. Lim recategorization)

Data References

US EPA. "Clean Watersheds Needs Survey 2012: Report to Congress," January 2016. https://www.epa.gov/cwns/clean-watersheds-needs-survey-cwns-2012-report-and-data.

USA Counties Data File Downloads. http://www.census.gov/support/USACdataDownloads.html#IPE



EPA Regions 1-10. Source: usepa.gov

I. CREATE METRICS TO CONTEXTUALIZE REPORTED NEED

Calculate the following metrics:

- Projected change in residential population receiving collection
- Projected percent change in residential population receiving collection
- Projected change in residential population receiving treatment
- Projected percent change in residential population receiving treatment

- Projected change in non-residential population receiving treatment
- Projected percent change in non-residential population receiving treatment
- Projected change in non-residential population receiving collection
- Projected percent change in non-residential population receiving collection
- Residential population density (using county area and population)
- Change and percent change in population density
- Percent change in median income

II. PROVIDE SUMMARY STATISTICS OF THE DATA

Provide summary tables and figures that describe the trends that you see in the data. You may include summary tables of the 10 best and worst performing MSAs in each resilience measure, density plots, histograms, box plots, or any other means to communicate the data to your reader.

- Top and Bottom 10 facilities with greatest/least need (perhaps separating them out by region)
- Tables with min, max, quartiles, median, and mean reported for relevant variables
- Boxplots, density plots and histograms of need by EPA region, climatic region (NOAA), state, or other grouping (**hint:** log transformations of data can help in identifying relationships in data when they appear to be highly skewed, non-normally distributed or otherwise difficult to differentiate.

Hint for R: for categorical data ("factor" data type in R), the `summary()` command will display frequencies, rather than the typical min, max, median, mean, and quantile summary.

Hint for R: To find specific facilities in this dataset (which has more than 8,000 entries), try using the `grep` command (https://stat.ethz.ch/R-manual/R-devel/library/base/html/grep.html) for string pattern matching. Setting the `ignore.case` option in the grep command to `TRUE` is particularly helpful since this dataset contains 'mixed case' entries (R is typically case-sensitive).

Example for finding 'Philadelphia' facilities:

```
> dat<-read.csv('cwns2012.csv')
> dat$FACILITY_NAME [grep("philadelphia", dat$FACILITY_CITY, ignore.case =
TRUE)]

[1] PHILADELPHIA POTW              Philadelphia (V) STP
[3] New Philadelphia WWTP & Sewers PHILADELPHIA WATER DEPT (NE)
[5] PHILADELPHIA WATER DEPT (SE)   PHILADELPHIA WATER DEPT (SW)

# (note that not all of these are in Philadelphia, PA. Apparently, there is
also a Philadelphia in MS, and one in NY). Look at the other columns for
these rows to see this.
```

III. TESTS OF ASSOCIATION

As a group, is there a significant difference in the magnitude of overall budgeted need for facilities with CSS versus facilities with MS4s (Municipal Separate Sewer Systems)? Use a t-test of difference in means to compare these two groups.

As a group, is there a significant difference between facilities' needs for those that are associated with a TMDL? Does this vary by region?

Is there evidence of significant regional differences in need? Use the chi-square test (`CrossTable()` function).

Choose five variables in the dataset, or from the variables that you created in Question I that you suspect may be associated with reported infrastructure need. Using the `CrossTable()`function to determine if there is evidence of association between each of these variables and total overall need. Where do the Philadelphia facilities fall amongst the categories that you created for each variable?

IV. CORRELATION

Choose five variables in the dataset, or from the variables that you created in Question I. Create scatterplots between these variables and reported need, and between each other. Use the `cor.test()` function to determine evidence of correlation between variables.

V. CONCLUSION

Draft a conclusion to the report to Mayor Kenney summarizing your findings. What additional variables would be helpful to include that might better explain the variation you observe if you had additional scope?