DataW Report

Amber O’Connell, aoconnel@ucsc.edu

CMPS 263 Winter 2016

March 18, 2017

https://github.com/amberatucsc/cmps-263-final-project

0. Tools used:

    a) Python 2.7, BeautifulSoup4 4.5, requests, codecs, re, pprint
    b) R 3.3.2, plyr 1.8, ggplot2 2.2, reshape2 1.4, tm 0.7, knitr 1.15
    c)MS excel for viewing csv’s

1.Data wrangling

The data, I found on the web as html/xml tables. I used the beautiful soup python library to translate the columns and rows of the tables into simple text entries that could be added to a plain text file (values separated by ‘;’ and columns new line separated).

I then translated the text files to csv’s and did some simple re-aligning/cleaning in MS excel. From there I loaded the csv’s into the rHTML code to turn the csv’s into dataframes. rHTML is a html document that you can inject r code into; it is provided by the r knitr package. (california_water.RHtml). I did some simple cleaning within R to get the county data from the water department to join with R’s California mapping data.

Inputs: https://cdec.water.ca.gov/cgi-progs/reservoirs/STORSUM
    http://cdec.water.ca.gov/misc/monthly_res.html
    http://www.california-demographics.com/counties_by_population

script: hw5.py

Outputs: reservoir_metadata.csv
    area_reservoir_totals_forr.csv
    california_population_est.csv

2. Data analysis

The water data while easy to get from the California water department, was not as easy to understand in context. The data was broken down in a few different geographical categories. The most common for the data I wanted was by hydrological region. As most map data in R or other graphing tools is by county or city or simple lat-lon, I manually mapped the county data to the most reasonable hydrologic region (thinking about county sizes, geographic features, etc).

Outputs: ca_county_hydrologicregion.csv

I also grabbed county population data to better understand water availability per capita.

3.Data visualization

The visualizations were completed in R. I used the ggplot package with R’s California mapping data. I created two plots of California county data, one with reservoir data and hydrologic region data and the other with population data. I also added two bar plots that look at water levels per hydrologic region and population data per hydrologic region.

Script: california_water.Rhtml

Output: california_water.html