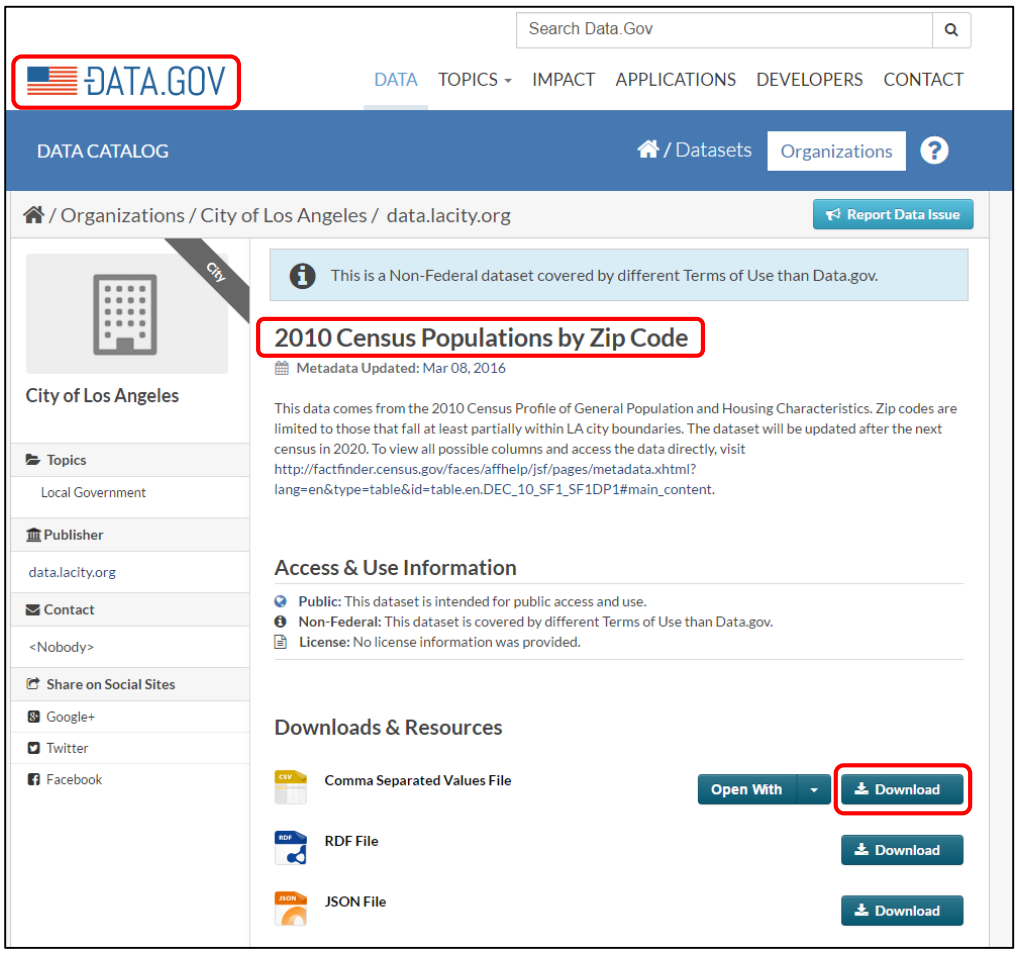
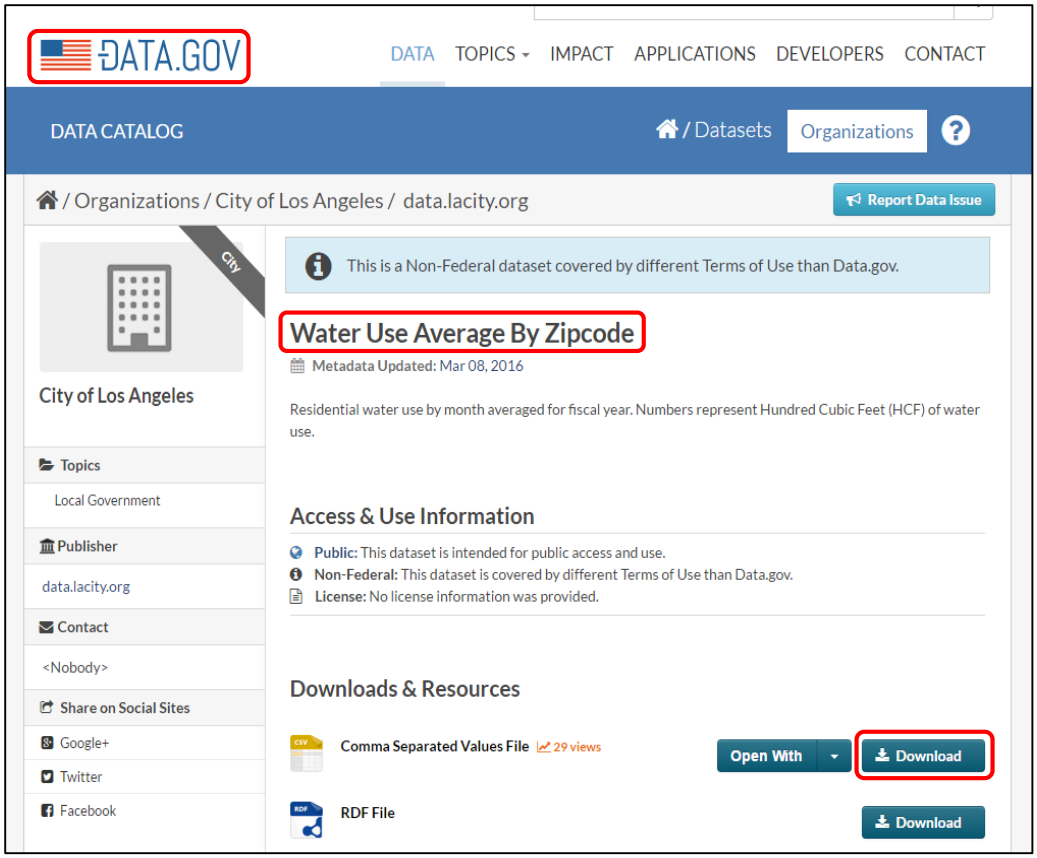
Workflow Documentation

The following steps indicate the data cleaning and merging process implemented.

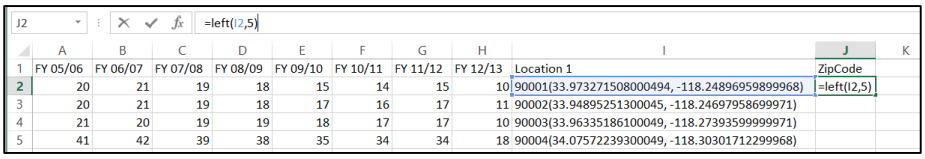
1. Go to the source of the focal dataset: http://catalog.data.gov/dataset/2010-census-populations-by-zip-code. Then download the dataset in csv (Comma Separated Values) format.



2. Go to the source of the first additional dataset (water usage): http://catalog.data.gov/dataset/water-use-average-by-zipcode-8dbe0. Then download the dataset in csv (Comma Separated Values) format.



3. Open the water usage dataset in MS Excel or similar program. The zip code in this dataset has location appended to it. So to match it with the focal dataset we have to remove the appended location.
    a. Create a new column “Zip Code”.
    b. Then use the command *=Left(I2,5)* this copy only the first 5 characters from location.



	A	B	C	D	E	F	G	H	I	J
1	FY 05/06	FY 06/07	FY 07/08	FY 08/09	FY 09/10	FY 10/11	FY 11/12	FY 12/13	Location 1	ZipCode
2	20	21	19	18	15	14	15	10	90001(33.973271508000494, -118.24896959899968)	=left(I2,5)
3	20	21	19	18	17	16	17	11	90002(33.94895251300045, -118.24697958699971)	
4	21	20	19	19	18	17	17	10	90003(33.96335186100049, -118.27393599999971)	
5	41	42	39	38	35	34	34	18	90004(34.07572239300049, -118.30301712299968)	

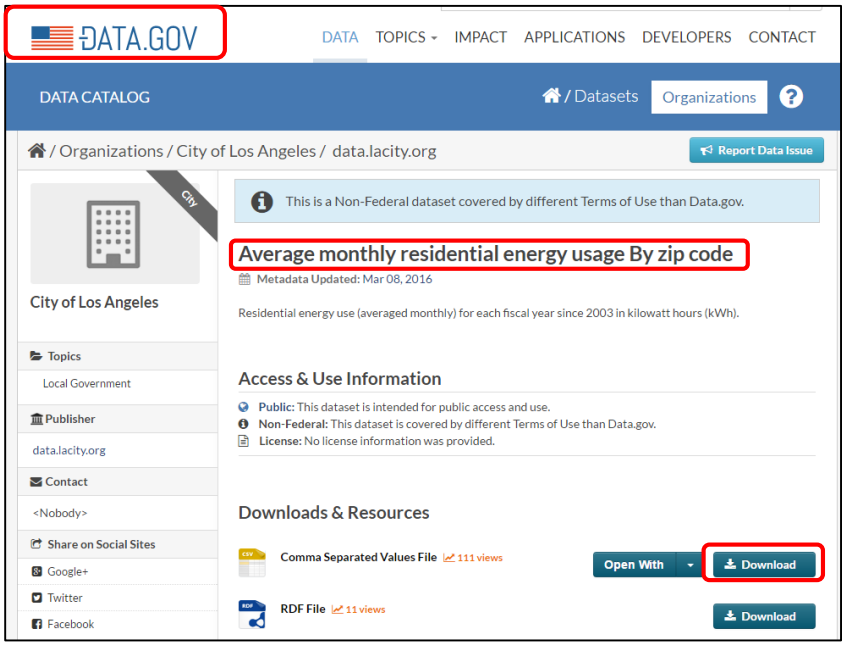
    c. Then select the cell and use the use the fill handle (double click on the small square on the right bottom) to repeat it for the remaining rows.

I	J
Location 1	ZipCode
90001(33.973271508000494, -118.24896959899968)	90001
90002(33.94895251300045, -118.24697958699971)	
90003(33.96335186100049, -118.27393599999971)	
90004(34.07572239300049, -118.30301712299968)	

I	J
Location 1	ZipCode
90001(33.973271508000494, -118.24896959899968)	90001
90002(33.94895251300045, -118.24697958699971)	90002
90003(33.96335186100049, -118.27393599999971)	90003
90004(34.07572239300049, -118.30301712299968)	90004

    d. Save the file in csv format.
    e. Reopen the file delete the *Location 1* column. Save the file in the csv format.

4. Go to the source of the first additional dataset (energy usage): http://catalog.data.gov/dataset/average-monthly-residential-energy-usage-by-zip-code-0487d. Then download the dataset in csv (Comma Separated Values) format.



5. Repeat the steps performed in Step 3 to extract zip code.
6. Run the R Script 'Merging Data.R' to merge the focal data with the additional datasets. This file uses the merge command to merge datasets. It first merges the 2 additional data sets and then merges the result to the focal dataset.

Limitations: The demographics data corresponds to the census of 2010. Since census is not taken every year the yearly increase or decrease in population could not be taken into consideration.