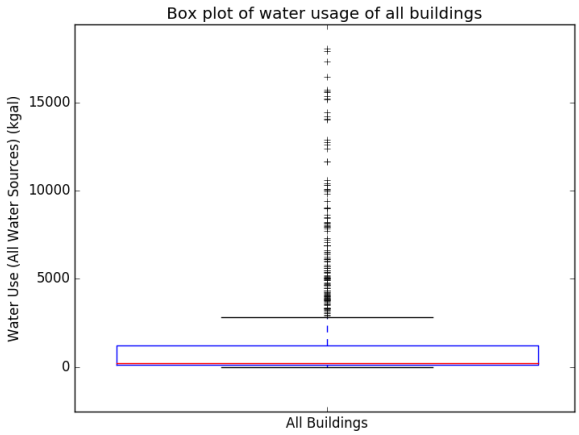
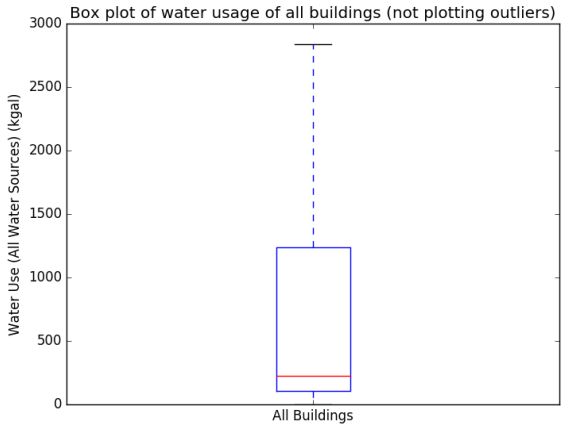
CS 5665 HW#1
Karun Joseph, A02240287

All code, plots reside here: https://github.com/karunmj/usu-coursework/tree/master/cs5660datasc/hw/hw1
Data preprocessing involved loading the csv file as a data frame object using Pandas library in Python

1. Water usage analysis:
   All NaN values in water usage column were replaced by the mean of available water usage data of other buildings.
   a. Water use for all buildings
      Mean with outliers - 6913.163341 kgal
      Median with outliers - 225.0 kgal
      Mode with outliers - 0 and 165.0 kgal

      Box plots for all building





From above plots and measures, it looks like there are a bunch of zero usage numbers. This is especially evident in the mode measure. There also seems to exist very high numbers that tend to pull the measures higher up.

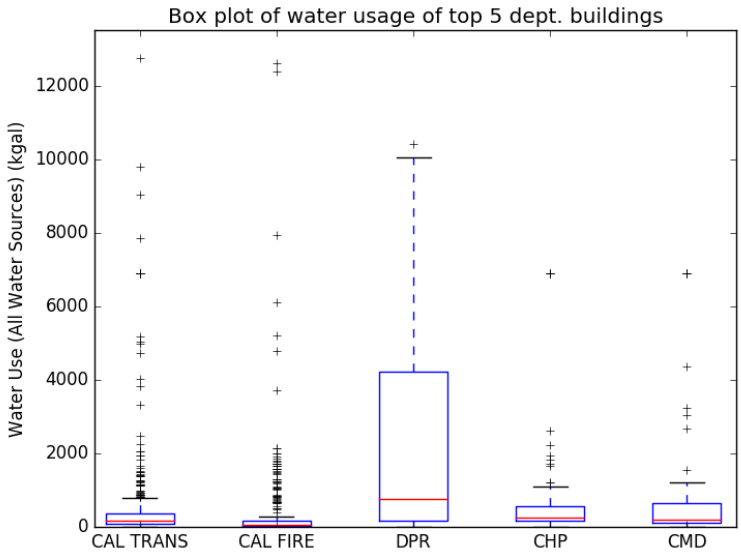
   b. Water use for top 5 department buildings
      The top 5 departments based on building numbers were found to be CAL TRANS, CAL FIRE, DPR, CHP and CMD. Their measure are as follows:

      CAL TRANS - Mean = 690.53, Median=165, Mode=165

CS 5665 HW#1
Karun Joseph, A02240287

CAL FIRE - Mean =501.88 , Median=58.6, Mode=51.5
DPR - Mean =2942.11 , Median=762.65, Mode=165
CHP - Mean = 1344.05, Median=256.7, Mode=0, 165
CMD - Mean =772.75 , Median=206.4, Mode=165

Box plot for Top 5 department buildings



From the plots and data, it looks like the top 5 depts. have similar behavior to all other buildings.

c. Removing outliers, water use for all buildings
All data that lies beyond (q1-1.5*(q3-q1)) and (q3+1.5*(q3-q1)), where q1 and q3 are first and third quartile have been removed as outliers.

Mean without outliers - 433.752430556 kgal
Median without outliers - 165.0 kgal
Mode without outliers – 165.0 kgal

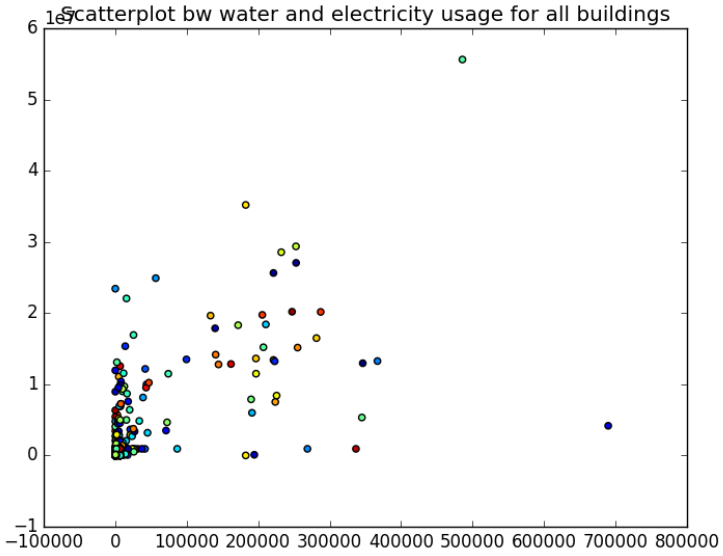
d. Removing outliers, water use for top 5 department buildings
CAL TRANS - Mean = 186.9, Median=165, Mode=165
CAL FIRE - Mean =64.35 , Median=51.5, Mode=51.5
DPR - Mean =2077.42, Median=700, Mode=165
CHP - Mean = 325.9, Median=235, Mode=165
CMD - Mean =324.24, Median=165, Mode=165

Removing outliers has made the data much cleaner, especially the zeroes and the very high values have been got rid of.

2. Resource usage correlation
   a. Between all buildings

CS 5665 HW#1
Karun Joseph, A02240287

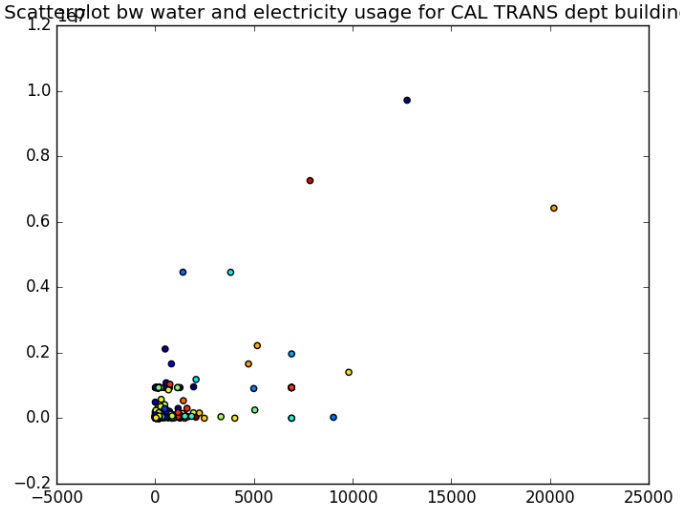
Scatter plot



Persons correlation bw electricity and water usage - 0.6661176605002519

Since the coefficient is greater than 0, both electricity and water usage has a positive correlation, i.e. as electricity usage increase so does water.

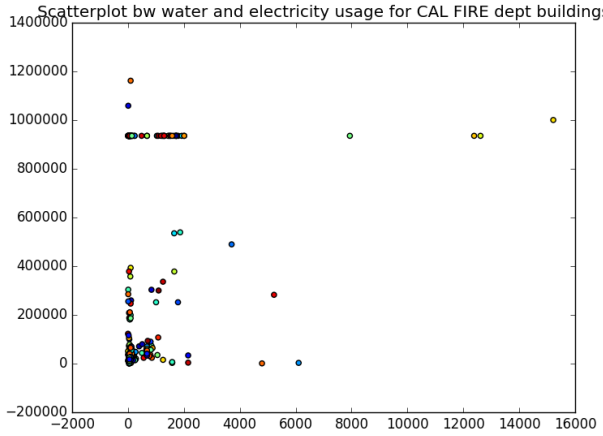
b. Between top 5 department buildings



Persons correlation bw electricity and water usage for CAL TRANS buildings - 0.59746992566881985

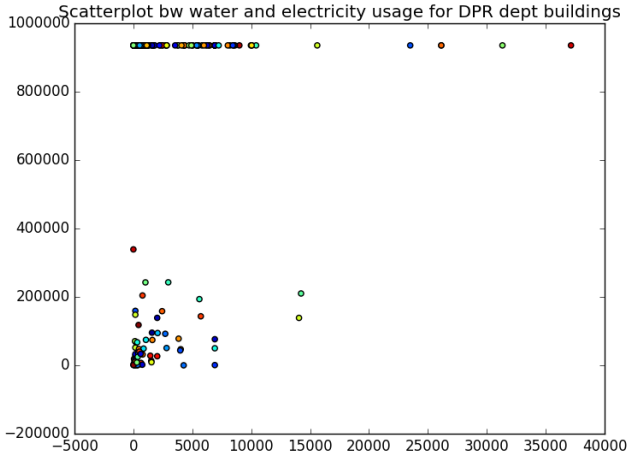
Since the coefficient is greater than 0, both electricity and water usage has a positive correlation, i.e. as electricity usage increase so does water.

CS 5665 HW#1
Karun Joseph, A02240287



Persons correlation bw electricity and water usage for CAL FIRE buildings - 0.24978554815196741

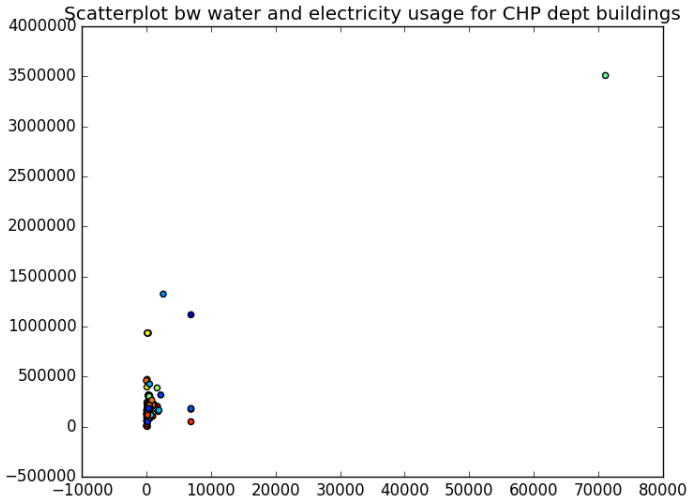
Since the coefficient is greater than 0, both electricity and water usage has a positive correlation, i.e. as electricity usage increase so does water. However, this value is very close to 0 which indicates there exists almost no correlation.



Persons correlation bw electricity and water usage for DPR buildings - 0.15380068291814197

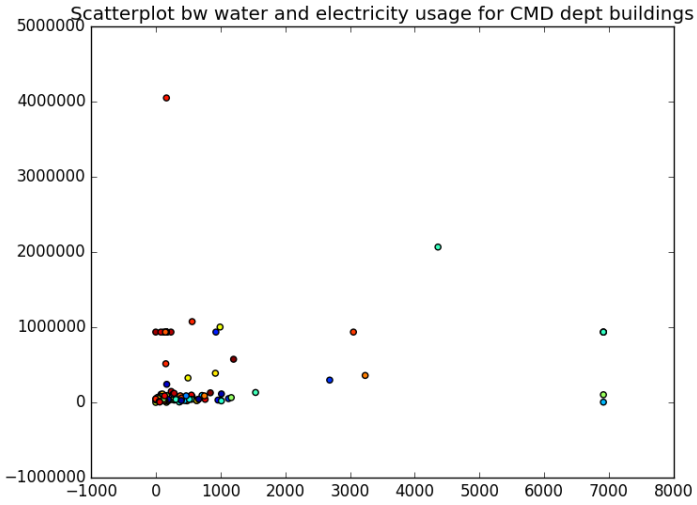
Since the coefficient is greater than 0, both electricity and water usage has a positive correlation, i.e. as electricity usage increase so does water. However, this value is very close to 0 which indicates there exists almost no correlation.

CS 5665 HW#1
Karun Joseph, A02240287



Persons correlation bw electricity and water usage for CHP buildings - 0.8167926077097335

Since the coefficient is greater than 0, both electricity and water usage has a positive correlation, i.e. as electricity usage increase so does water. However, this value is very close to 1 which indicates there exists a likelihood of linear relationship.



Persons correlation bw electricity and water usage for CMD buildings - 0.21429228623294547

Since the coefficient is greater than 0, both electricity and water usage has a positive correlation, i.e. as electricity usage increase so does water. However, this value is very close to 0 which indicates there exists almost no correlation.

Overall, there exists a positive correlation between electricity and water usage

3. Building similarities
    a. Resource usage

All the variables are of type quantitative. Just like in the first answer, any missing data is replaced with the mean of existing data.

CS 5665 HW#1
Karun Joseph, A02240287

	Manhattan		
	First	Second	Third
Mendota maintenance st.	OROVILLE AREA (5710.5)	Torrance (State Owned) (12913.64)	Orange (State Owned) (15479.04)
Metropolitan state hospital	DAA 22, SAN DIEGO COUNTY FAIRGROUNDS (667634.1)	PATTON STATE HOSPITAL (3600558.56)	MEADOWVIEW (3984573)
Long beach field office	CSR-SLU San Luis Obispo FS - 2014 E Complete *	AMERICAN RIVER FISH HATCHERY *	CAJON MAINTENANCE STATION *

	Euclidean		
	First	Second	Third
Mendota maintenance st.	OROVILLE AREA (3931.75)	Torrance (State Owned) (10262.35)	FREMONT MAINTENANCE STATION (13123.36)
Metropolitan state hospital	DAA 22, SAN DIEGO COUNTY FAIRGROUNDS (578653.06)	PATTON STATE HOSPITAL (3593816.7)	MEADOWVIEW (3969122.5)
Long beach field office	CSR-SLU San Luis Obispo FS - 2014 E Complete *	AMERICAN RIVER FISH HATCHERY *	925 BOLSA CHICA SB *

	Cosine		
	First	Second	Third
Mendota maintenance st.	OROVILLE AREA (7.18e-5)	Torrance (State Owned) (0.0004)	FERRELLGAS (0.0007)
Metropolitan state hospital	DAA 22, SAN DIEGO COUNTY FAIRGROUNDS (0.0005)	SOUTHERN DIVISION HEADQUARTERS (0.009)	PATTON STATE HOSPITAL (0.013)
Long beach field office	CSR-SLU San Luis Obispo FS - 2014 E Complete (3.44e-6)	AMERICAN RIVER FISH HATCHERY *	CAJON MAINTENANCE STATION*

b. Property variables

Since the quantities involved are nominal (except for property area), these are first converted to quantitative. I have assigned a unique integer to each string element in the column. These strings are also unique (I have preprocessed the column to regard both lower and upper cases of a text as same). For example, the 'City' attribute has 642 unique entries, and each of them has been assigned an integer index to calculate similarity metrics.

CS 5665 HW#1
Karun Joseph, A02240287

	Manhattan		
	First	Second	Third
Mendota maintenance st.	ALAMEDA MAINTENANCE STATION*	SKYLONDA STORAGE *	WASCO MAINTENANCE STATION *
Metropolitan state hospital	PBSP-PELICAN BAY STATE PRISON *	LAC- CALIFORNIA STATE PRISON, LOS ANGELES COUNTY *	Sonoma DC *
Long beach field office	Chula Vista Maintenance Station *	MONTEBELLO OFFICE BUILDING *	715 CASTLE ROCK SP *

	Euclidean		
	First	Second	Third
Mendota maintenance st.	ALAMEDA MAINTENANCE STATION *	SKYLONDA STORAGE *	WASCO MAINTENANCE STATION *
Metropolitan state hospital	PBSP-PELICAN BAY STATE PRISON *	LAC- CALIFORNIA STATE PRISON, LOS ANGELES COUNTY *	Sonoma DC *
Long beach field office	Chula Vista Maintenance Station *	715 CASTLE ROCK SP *	MONTEBELLO OFFICE BUILDING *

	Cosine		
	First	Second	Third
Mendota maintenance st.	1510 O ST DON CARLOS APARTMENTS	BIG SYCAMORE MAINTENANCE STATION	17TH STREET COMMONS/MIXED USE
Metropolitan state hospital	DAA 22, SAN DIEGO COUNTY FAIRGROUNDS	CIM-CALIFORNIA INSTITUTION FOR MEN	KVSP-KERN VALLEY STATE PRISON
Long beach field office	BELL GARDENS OFFICE BUILDING	Pomona (Park) Armory (State Owned) (Asset Mana...	BUTTONWILLOW AREA

c. Resource usage and property variables

CS 5665 HW#1
Karun Joseph, A02240287

	Manhattan		
	First	Second	Third
Mendota maintenance st.	OROVILLE AREA	FREMONT MAINTENANCE STATION	MANZANITA MAINTENANCE STATION
Metropolitan state hospital	DAA 22, SAN DIEGO COUNTY FAIRGROUND	PATTON STATE HOSPITAL	DMV HQ Campus - East Building
Long beach field office	AMERICAN RIVER FISH HATCHERY	CAJON MAINTENANCE STATION	925 BOLSA CHICA SB

	Euclidean		
	First	Second	Third
Mendota maintenance st.	OROVILLE AREA	FREMONT MAINTENANCE STATION	MANZANITA MAINTENANCE STATION
Metropolitan state hospital	DAA 22, SAN DIEGO COUNTY FAIRGROUND	PATTON STATE HOSPITAL	MEADOWVIEW
Long beach field office	AMERICAN RIVER FISH HATCHERY	CAJON MAINTENANCE STATION	925 BOLSA CHICA SB

	Cosine		
	First	Second	Third
Mendota maintenance st.	OROVILLE AREA	FERRELLGAS	MANZANITA MAINTENANCE STATION
Metropolitan state hospital	DAA 22, SAN DIEGO COUNTY FAIRGROUND	SOUTHERN DIVISION HEADQUARTERS	PATTON STATE HOSPITAL
Long beach field office	AMERICAN RIVER FISH HATCHERY	CAJON MAINTENANCE STATION	BISHOP AREA

(* Exact distance numbers can be obtained from script, however not all of them are included in this report)