Homework #1

CS 5665, Fall 2016

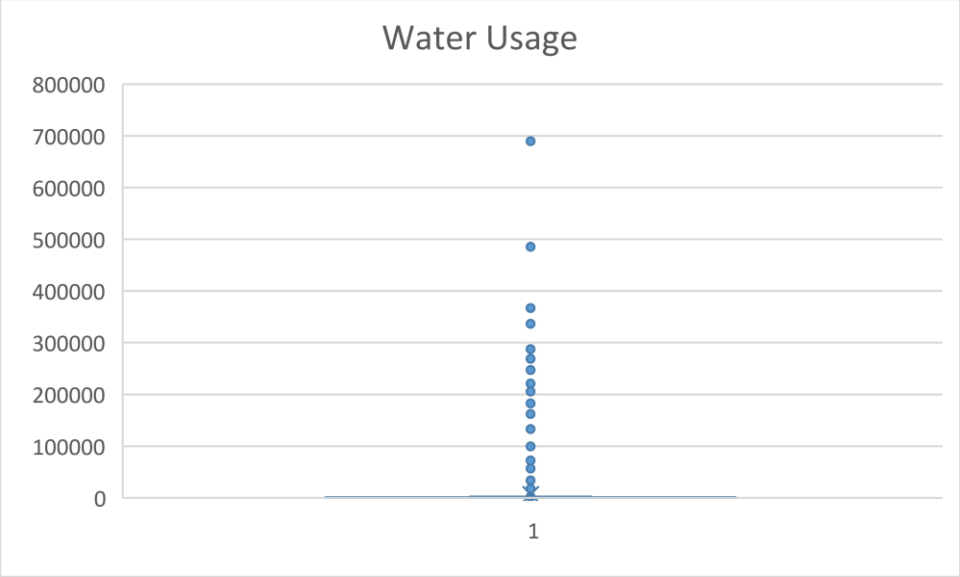
For assignment #1 I have used following stuffs:

- For this assignment I have used Microsoft Excel 2013 for data manipulation and visualization
- I have taken the relevant data from the csv files by copying and pasting them into different sheets created for different tasks.
- For handling missing attributes, I have not made any changes to data.

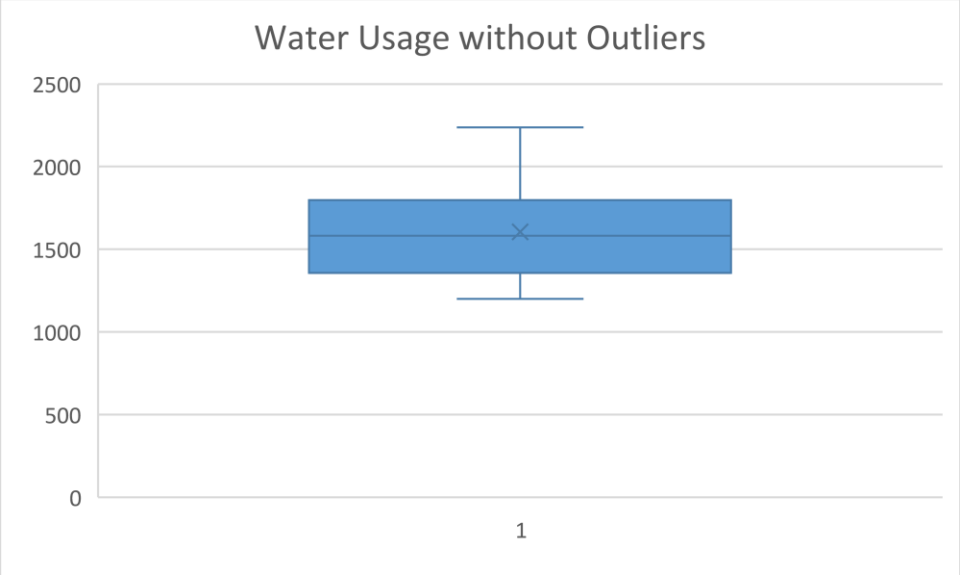
Task 1. Water Usage Analysis: First, we want to gain an understanding of the Water Usage of all buildings, and how these may vary across different departments.

(a) Using the three main measures of central tendency (mean, median and mode), analyze the Water Use for all buildings, as well as for individual departments (say, for the top-5 departments). You should plot box-plots for all buildings, as well as for the individual departments.

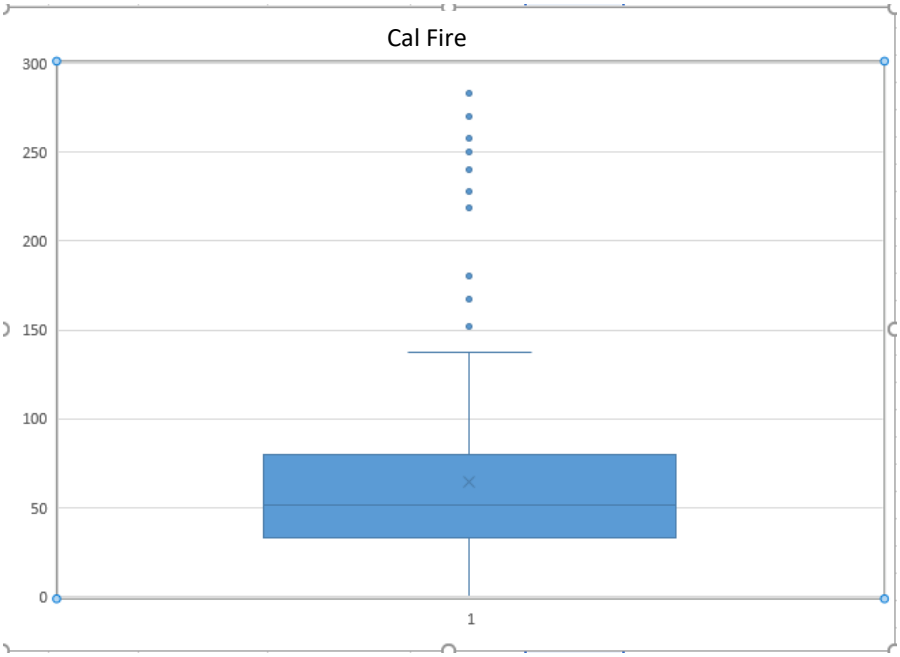
Ans:

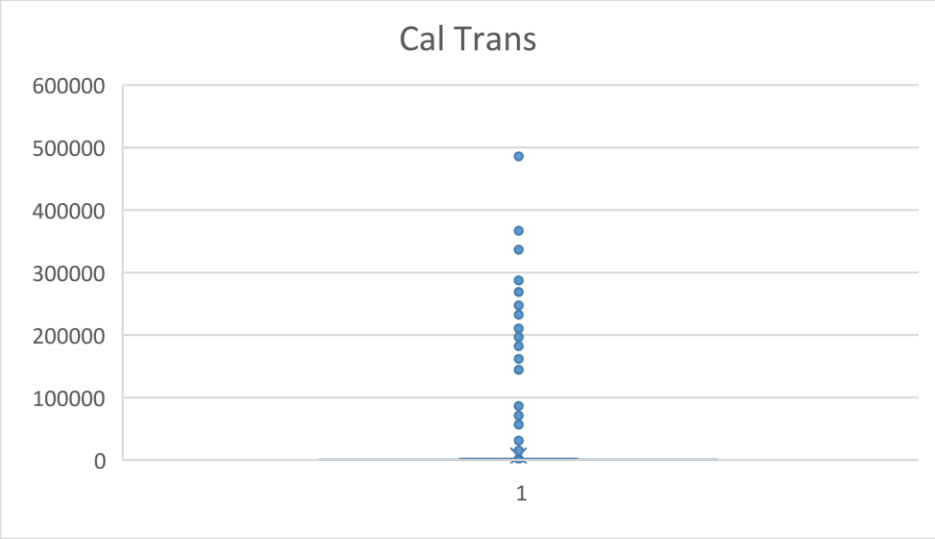


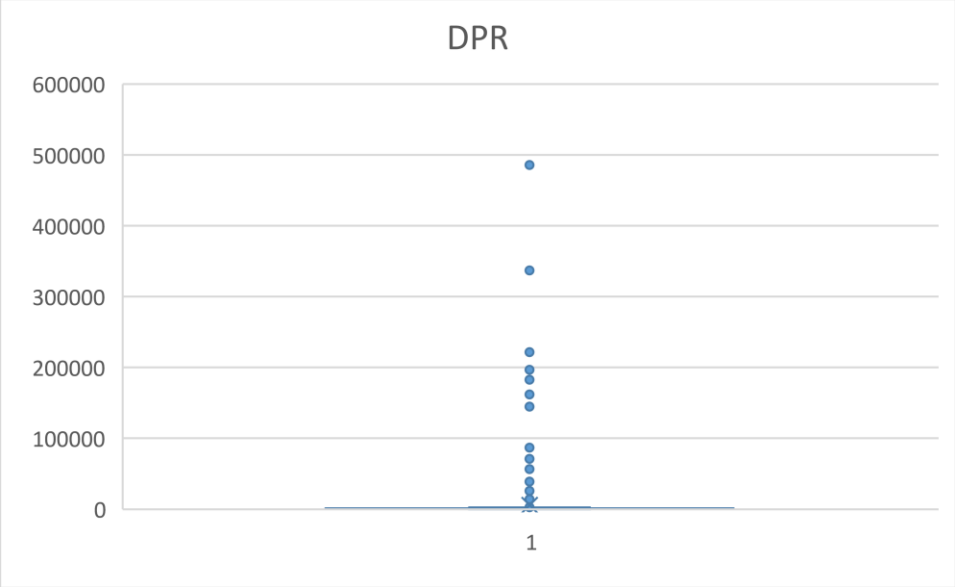
s

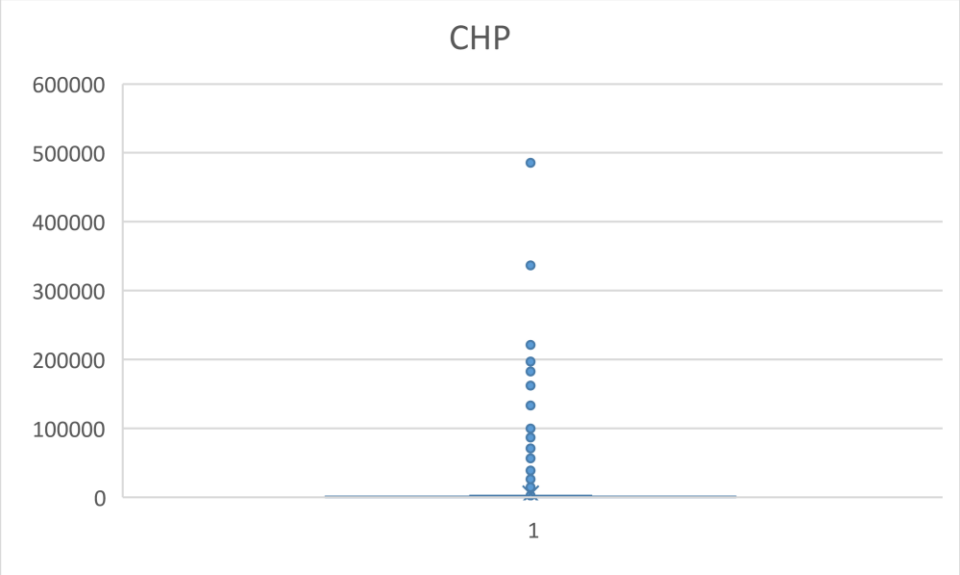


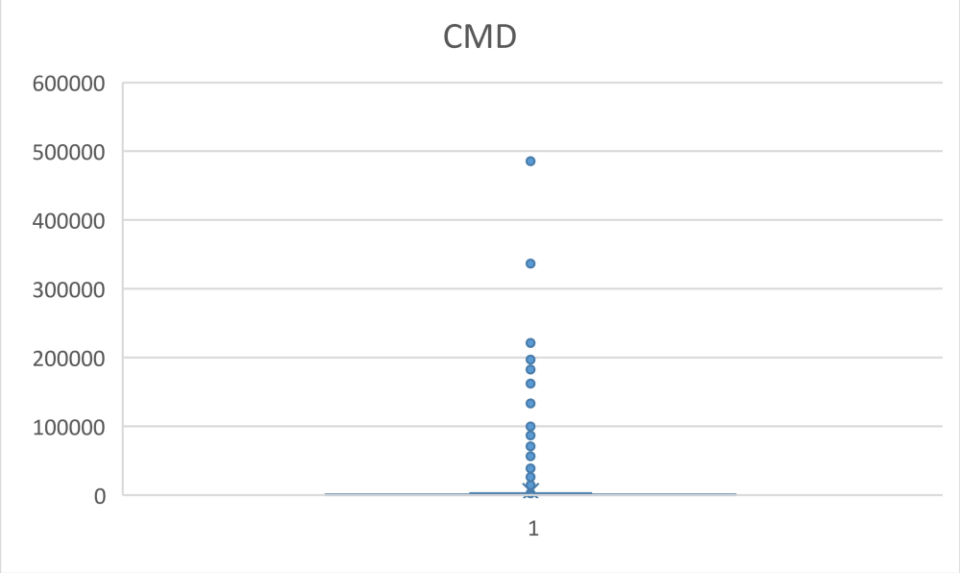
Top 5 departments:




Cal Trans
600000
500000
400000
300000
200000
100000
0
1


DPR
600000
500000
400000
300000
200000
100000
0
1





Normal Mean	Normal Median	Normal Mode	Quartile 1	Quartile 3	Interquartile Range	Lower Bound	Upper Bound
6563.689076	224.6	165	96.075	952.825	856.75	1189.05	2237.95

(b) Now, remove outliers by dropping all water usage that are "too extreme". Be sure to quantify your definition of "too extreme" and explain how you arrived at that definition. Compare the mean, median, and mode without outliers. What do you observe?

Ans:

Too extreme/outlier: These are the values that are obtained by removing most extreme values. But for me I have calculated "Lower Bound & Upper Bound" value using the Interquartile value and then removed the values which are lesser than lower bound and greater than upper bound values.

Outlier Mean	Outlier Median	Outlier Mode
1618.2224	1580.8	1256

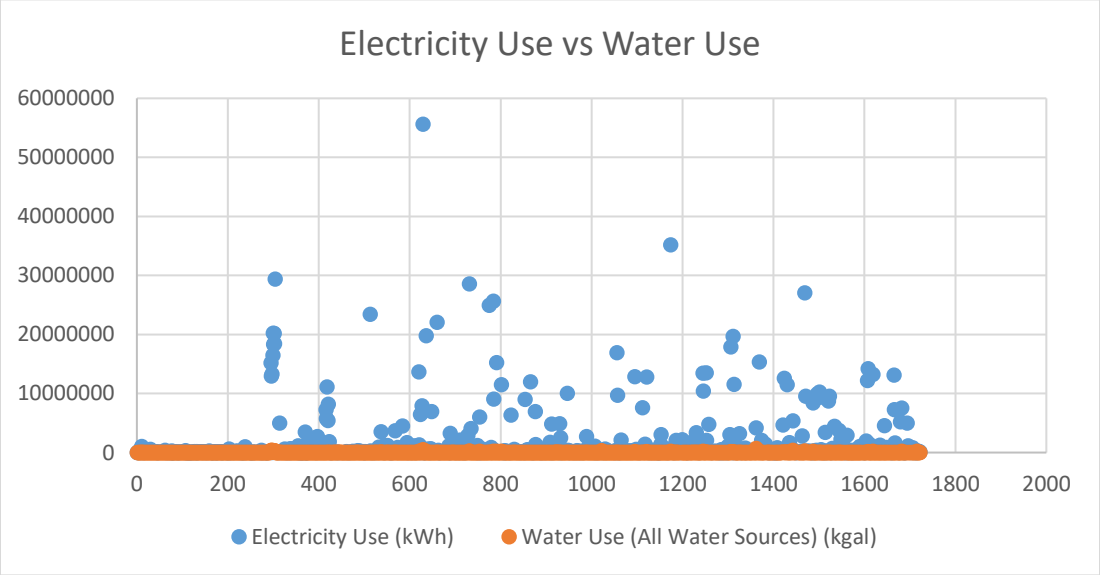
The central tendency values have changed very much. The mean value has decreased while the median and the mode value increased a lot. But theoretically, the mean value gets affected the most while median and mode change very less.

Task 2. Resource Usage Correlation: Does a relationship hold between the Water Use of a Building and its Electricity Use?

(a) Plot this relationship using a scatter plot, and report the correlation (using Person's correlation coefficient).

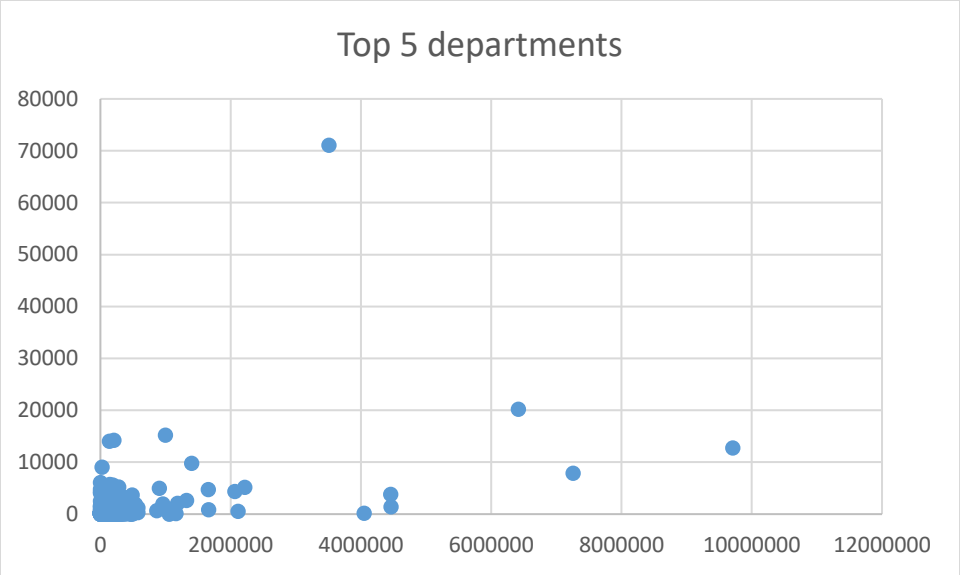
Ans:

Pearson's Correlation
0.700323281



(b) Now find the top 5 departments based upon the number of buildings and perform the same analysis for these 5 departments. Based upon these plots and Pearson correlation values what can you conclude?

Ans:



The Pearson coefficient of top 5 department:

Cal Fire - 0.299

Cal Trans - 0.732

CMD – 0.286

DPR – 0.108

CHP – 0.897

Task 3. Building Similarities: Using two distance metrics (Euclidean and Manhattan) and one similarity function (Cosine), find the three buildings similar to following buildings.

Property Name
MENDOTA MAINTENANCE STATION
METROPOLITAN STATE HOSPITAL
LONG BEACH FIELD OFFICE

(a) Resource Usage only: Electricity Use, Natural Gas Use, Propane Use, Water Use, Site Energy Use
(b) Property variables only : Department Name,City, Primary Property Type, Property Area
(c) All above 2 types of dimensions together

Ans: The similar buildings using the dimensions provided in the question are as follows:

MENDOTA MAINTENANCE STATION			
	Euclidean Distance	Manhattan distance	Cosine similarity

Resource Usage Only	- Oroville area 3931.757728 - Torrance (State Owned) 10262.35744 - Fremont maintenance station 13123.35676	- Oroville area 5710.499996 - Torrance (State Owned) 12913.6378 - Orange (State Owned) 15479.04278	- Oroville area 0.0000717690 - Torrance (State Owned) 0.000434664 - Ferrellgas 0.000753166
Property Variables Only	-Alameda Maintenance Station 29.0688837074973 -Mount Shasta Area 31.559467676119 -Vincent S/S 73.6613874428116	-Alameda Maintenance Station 31 -Mount Shasta Area 48 -Vincent S/S 86	-Vincent thomas bridge maintenance station 0.0000030580135181868 (Paint) -Porterville area 0.00000349995732 -Willows area 0.00000412088910
All dimensions	- Oroville area 4325.554511 - Fremont maintenance station 13124.97823 - Manzanita maintenance station 13961.42949	- Oroville area 7583.499996 - Fremont maintenance station 15965.19999 - Manzanita maintenance station 20422.64581	- Oroville area 0.0000939397 - Ferrellgas 0.000804833 - Manzanita maintenance station 0.000931998

METROPOLITAN STATE HOSPITAL			
	Euclidean Distance	Manhattan distance	Cosine similarity
Resource Usage Only	- Daa 22, San diego county fairgrounds 578653.0492 - Patton state hospital 3593816.735 - Meadowview	- Daa 22, San diego county fairgrounds 667634.0056 - Patton state hospital 3600558.4 - Meadowview	- Daa 22, San diego county fairgrounds 0.000494473 - Southern division headquarters 0.00901646 - Patton state hospital

	3969122.589	3984573	0.013671387
Property Variables Only	- PBSP-Pelican bay state prison 2556.72212 - Lac- California state 10538.46398 prison, los angeles county - Sonoma DC 20667.69532	- PBSP-Pelican bay state prison 2820 - Lac- California state 10665prison, los angeles county - Sonoma DC 20854	- Patton state hospital 0.0000000002 - Sol-California state prison, Solano 0.0000000003 - SQ-san quentin state prison 0.0000000009
All dimensions	- Daa 22, San diego county fairgrounds 695544.4996 - Patton state hospital 3593896.719 - Meadowview 4116497.937	- Daa 22, San diego county fairgrounds 1053804.096 - Patton state hospital 3624571.563 - DMV HW campus – east building 4805542.313	- Daa 22, San diego county fairgrounds 0.000831958 - Southern division headquarters 0.013683101 - Patton state hospital 0.018510132

LONG BEACH FIELD OFFICE			
	Euclidean Distance	Manhattan distance	Cosine similarity
Resource Usage Only	- Csr-Slu San Luis Obispo FS - 2014 E Complete 3299.738234 - American river fish hatchery 4380.921702 - 925 Bolsa chica sb 19283.06237	- Csr-Slu San Luis Obispo FS - 2014 E Complete 5188.3 - American river fish hatchery 6810.4 - Cajon maintenance station 25727.9	- Csr-Slu San Luis Obispo FS - 2014 E Complete 0.0000034430 - American river fish hatchery 0.0000127924 - Cajon maintenance station 0.0000392245
Property Variables Only	- 715 Castle rock SP 52.1056618804521 - Montebello office building 104.6565813 - Oroville (State Owned) 139.8320421	- 715 Castle rock SP - 57 - Montebello office building - 141 - Crescent city maintenance system -196	- 715 Castle rock SP 0.00000001489 - 730 Turlock lake SRA 0.00000008743 - Oakland coliseum building 0.00000019482

All dimensions	- American river fish hatchery 5841.24661 - Cajon maintenance station 19657.53805 - 925 Bolsa chica sb 29169.86024	- American river fish hatchery 10835.4 - Cajon maintenance station 26997.9 - 925 Bolsa chica sb 49305.7	- American river fish hatchery 0.00002771619 - Cajon maintenance station 0.00004117034 - Bishop Area 0.00015215278