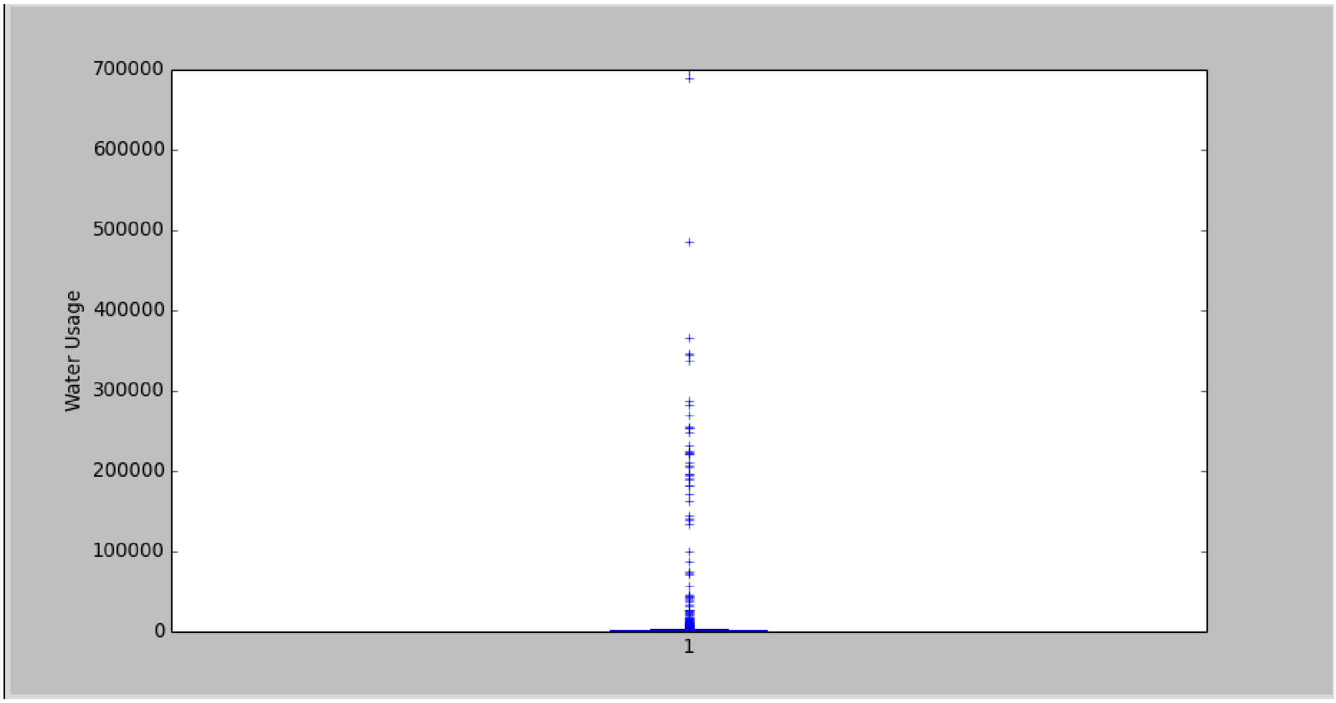
Homework #1
CS 5665, Fall 2016

1. Describe data transformation, issues faced and how you resolved ?
Below mentioned is common over all task.
+ Transformation:
- For loading used *Pandas* lib from python.
- For performing numerical operations *numpy* lib in python.
- For plotting figures (boxplot, scatter plot) used *matplotlib*.
- For performing scientific calculation (distance metrics, Cosine similarity) used *SciPy*

+ Issues:
- Blank values/NaN data/Empty strings: Calculated the mean of data and replaced with mean.
- Zero for electricity or water usage for buildings: Calculated mean of data and replaced Zero's with mean.
- Mean value seems more logical to replace NaN/Zero values.

2. Water Usage Analysis:

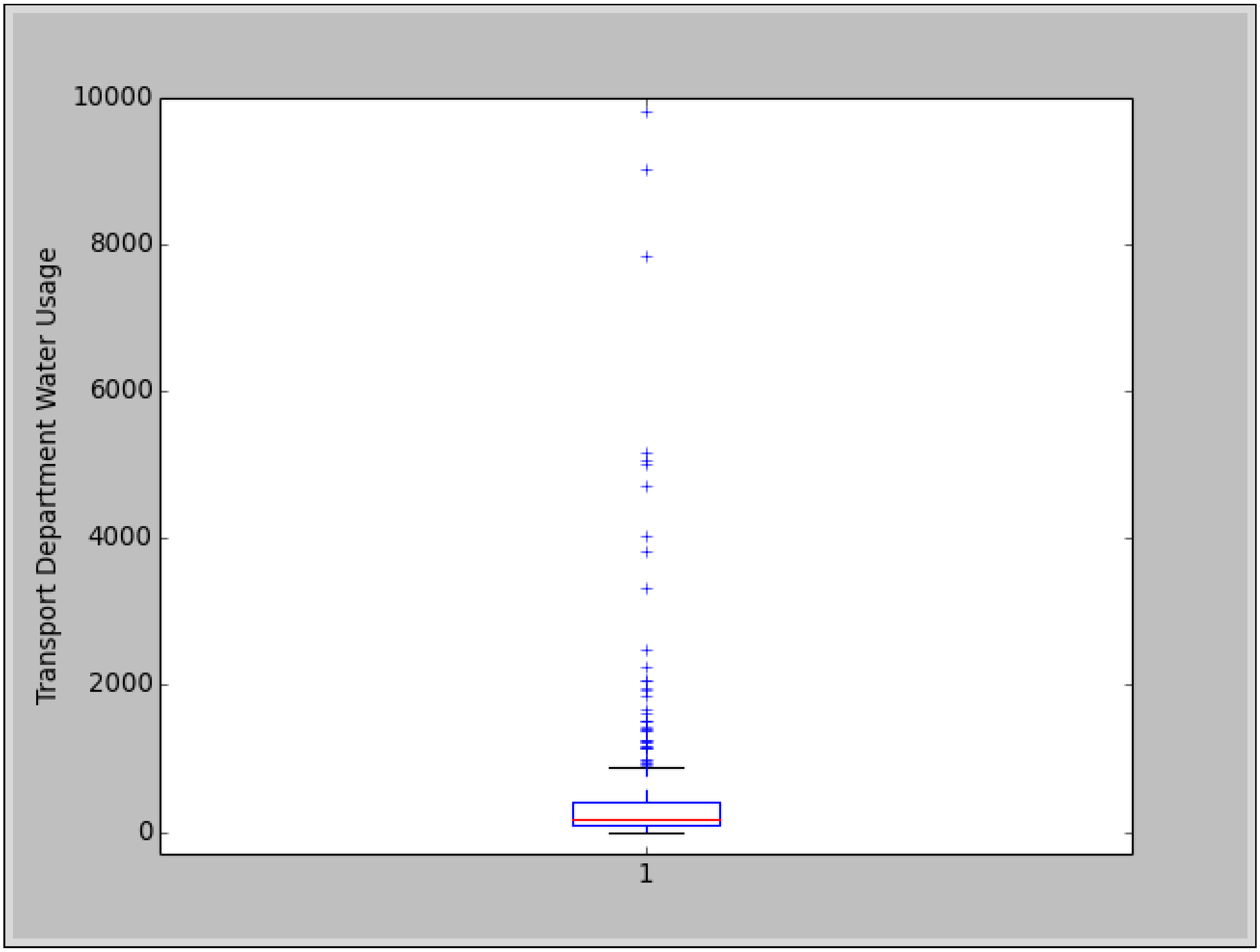
Box plot for Water Usage for all buildings:



Mean = 6989.44077816
Median = 227.95
Mode = 165.0

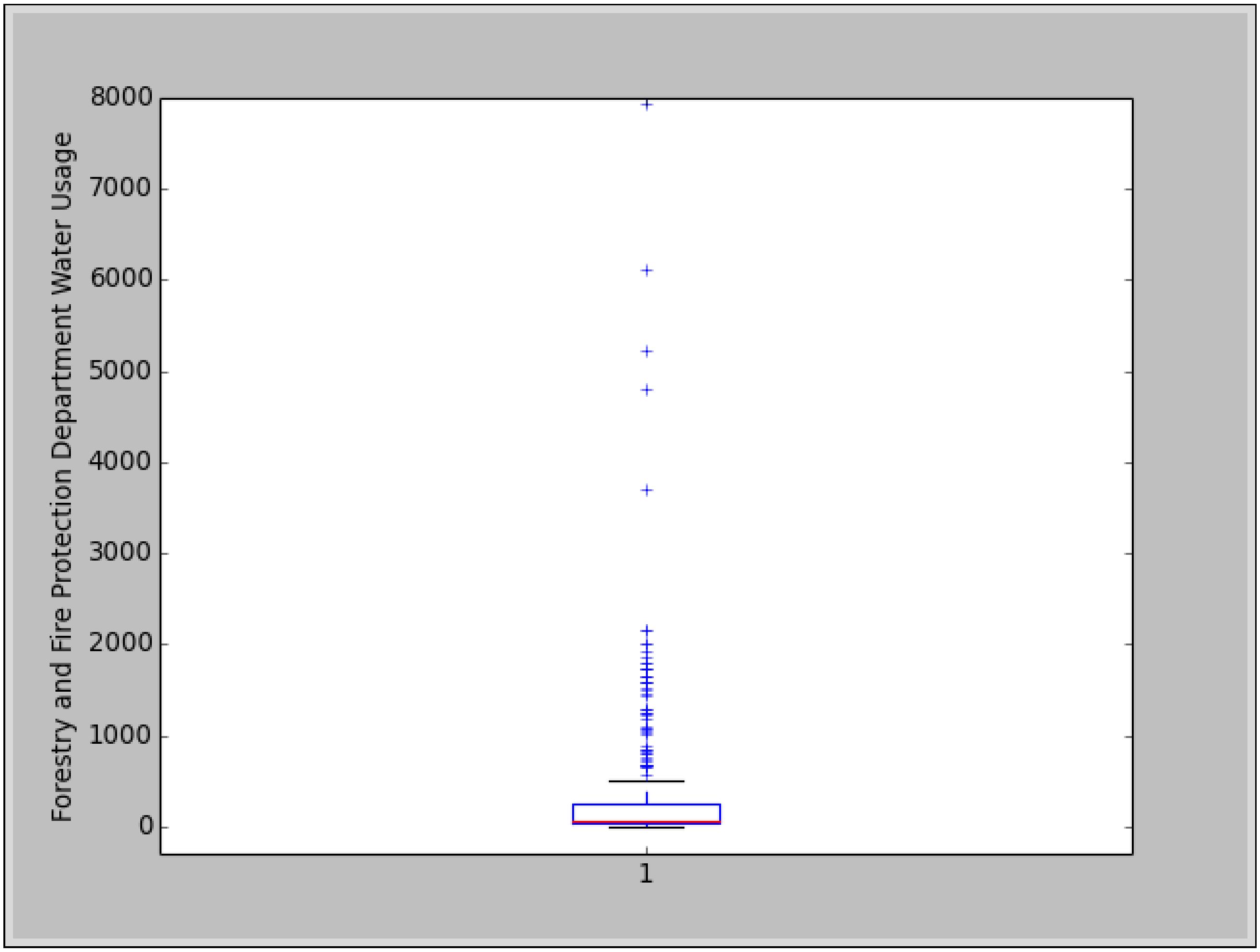
# Top 5 departments
# California Department of Transportation : 443
# California Department of Forestry and Fire Protection: 313
# Department of Parks and Recreation: 208
# California Highway Patrol: 107
# California Military Department: 103

California Department of Transportation



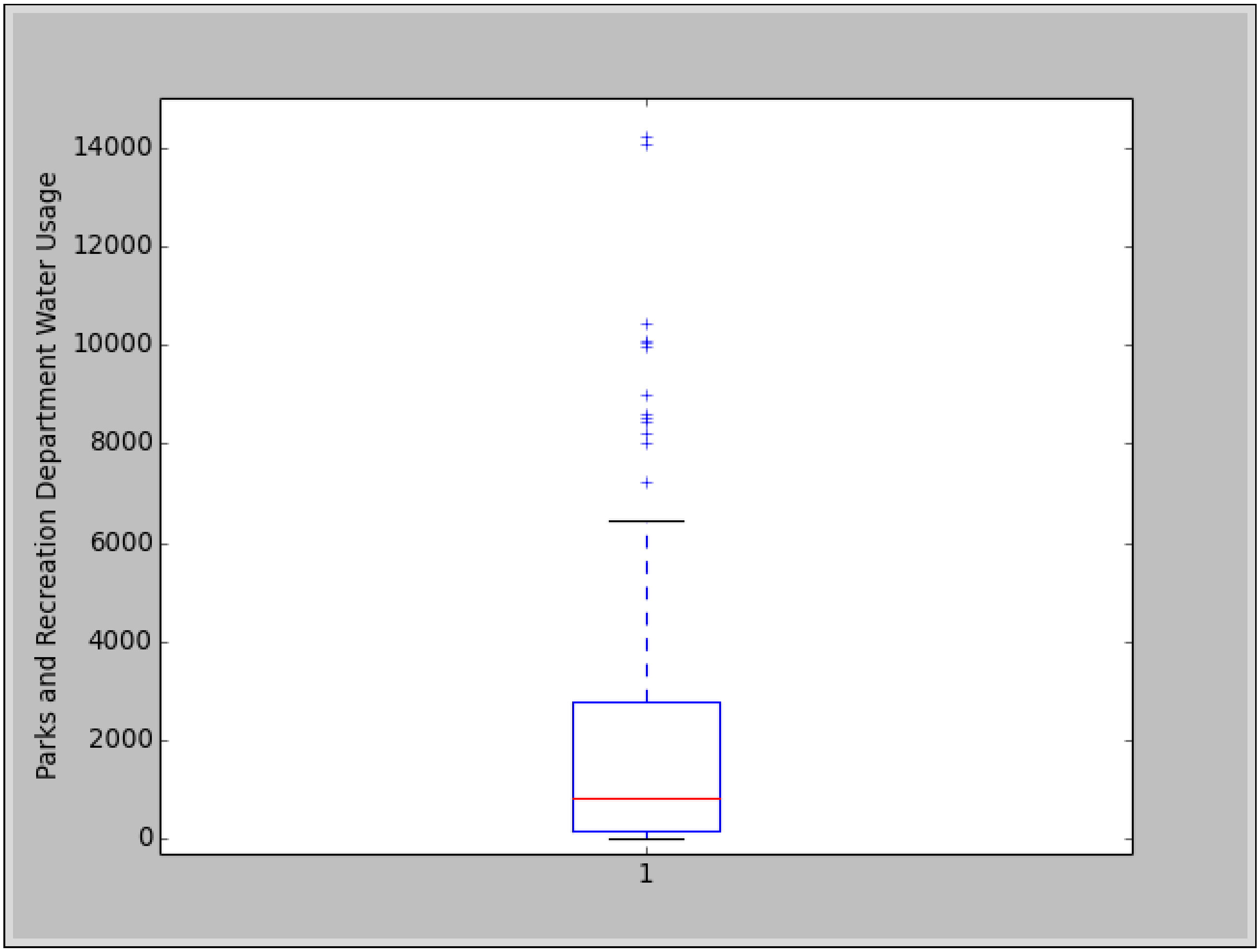
Mean = 508.076543651
Median =165.0
Mode = 165.0

California Department of Forestry and Fire Protection



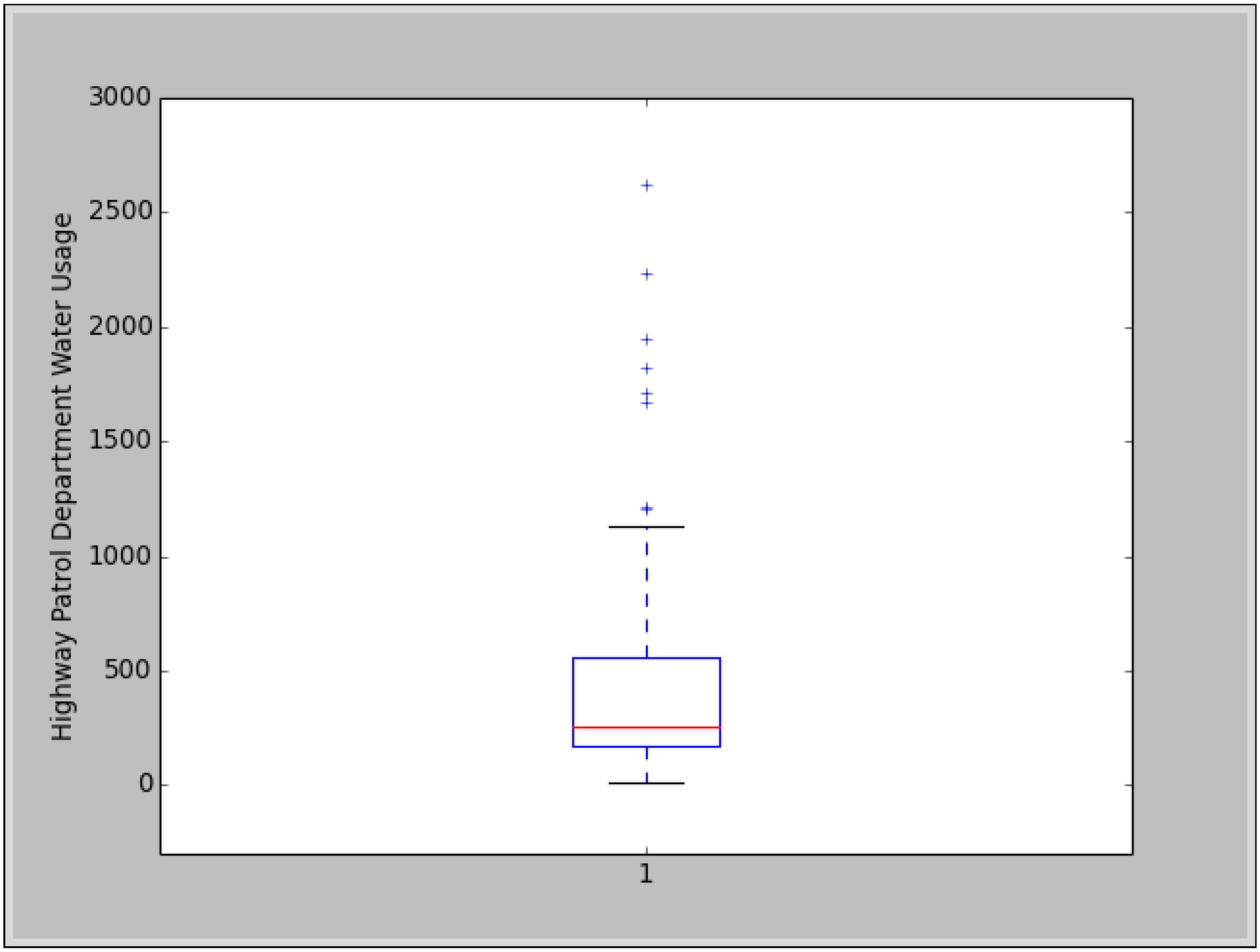
Mean = 508.295613919
Median = 63.3
Mode = 51.5

Department of Parks and Recreation



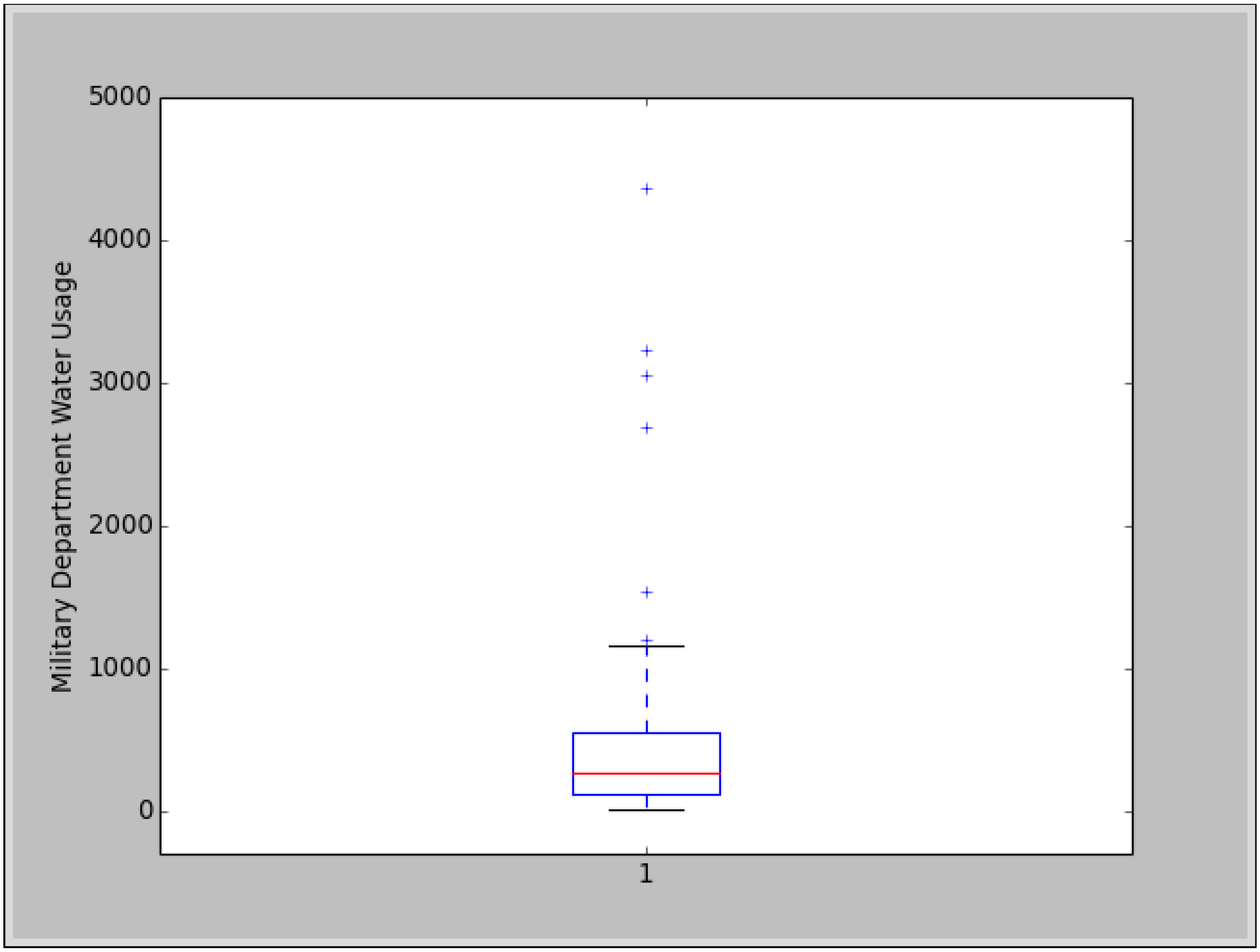
Mean = 2737.91080504
Median = 845.3
Mode = 165.0

California Highway Patrol



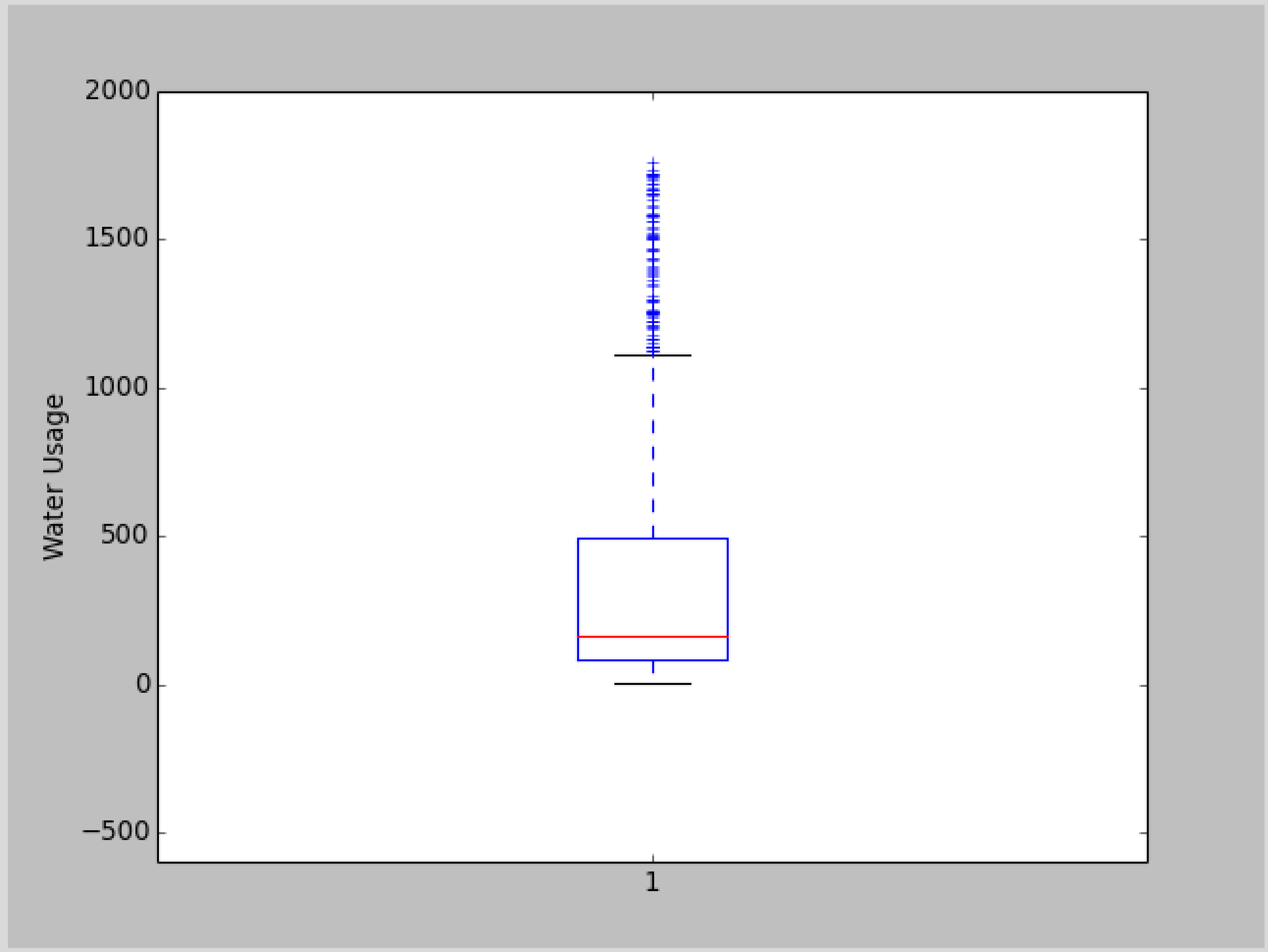
Mean = 1127.77669903
Median = 256.7
Mode = 165.0

California Military Department



Mean = 481.769447196
Median = 271.9
Mode = 459.465306122

Water Usage Box plot without outliers.



In this case, outliers are anything beyond ±1.5*IQR (InterQuartile Range)

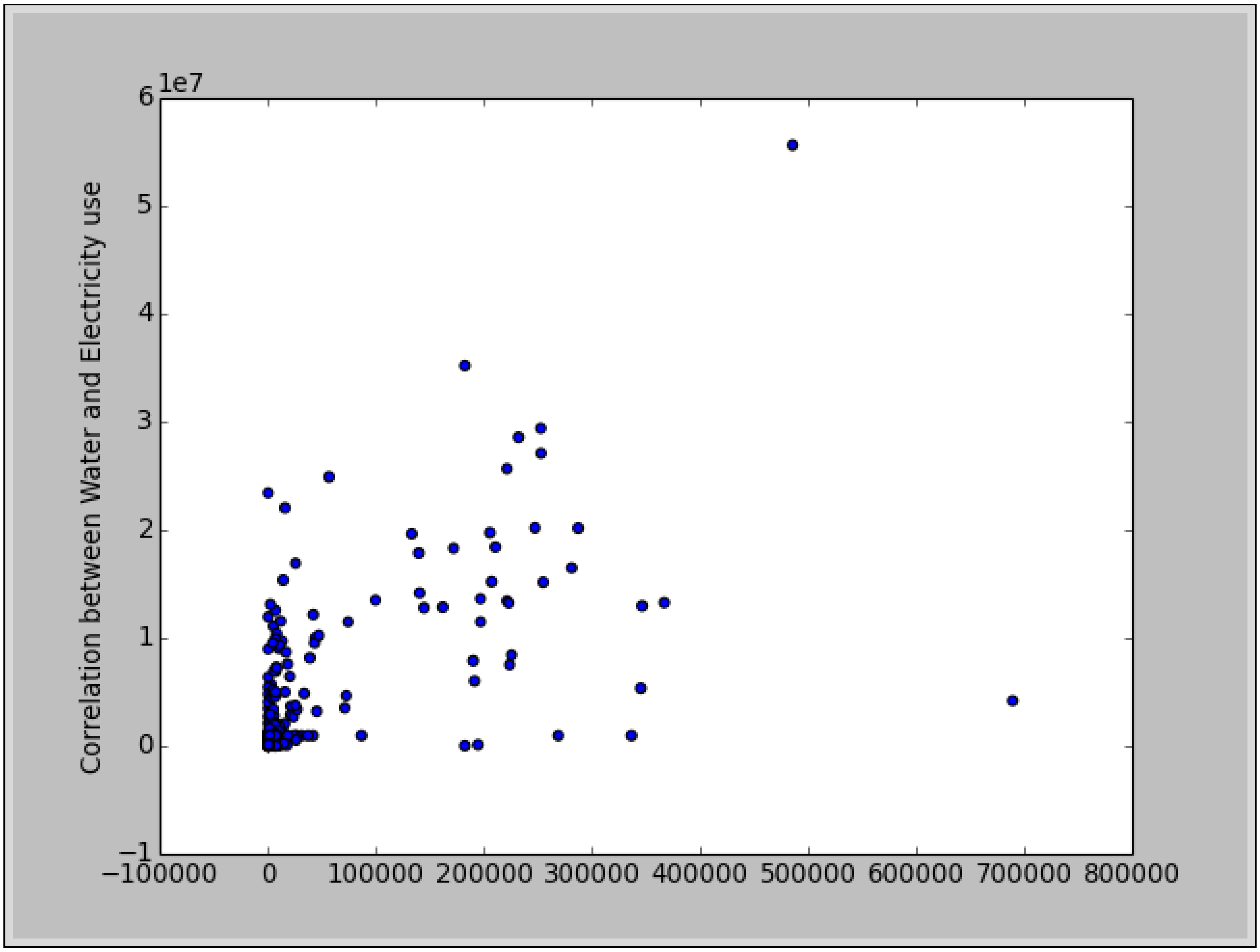
Mean, Median, Mode for data without outliers.
Mean = 350.485376662
Median = 165.0
Mode = 165.0

With outliers
Mean = 6989.44077816
Median = 227.95
Mode = 165.0

Measures of central tendency with removal of outliers are more stable and box plot also reveals it.

2. **Resource Usage Correlation:**

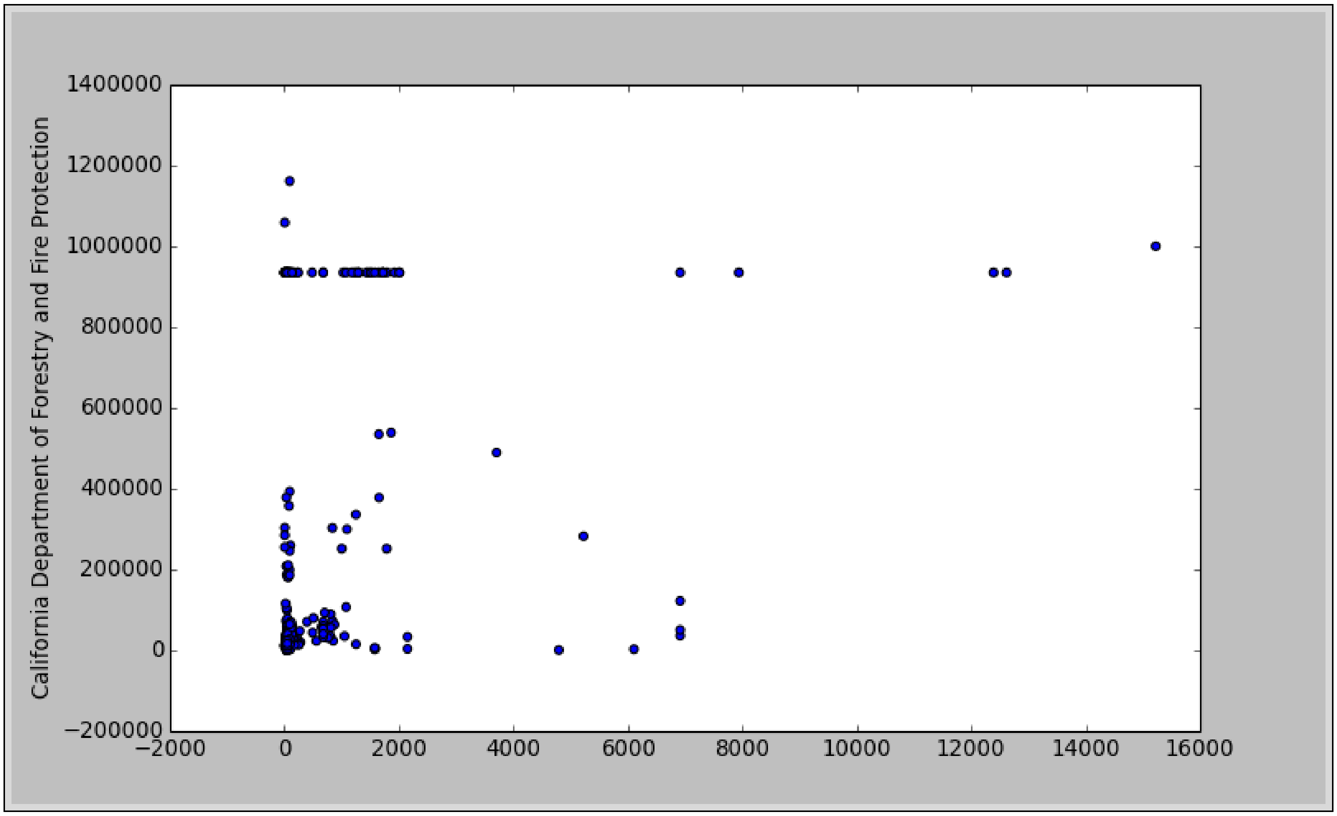
Relation between Water use and Electricity use of buildings
Pearson Correlation
(0.66577579621732319)



```
# Top 5 departments
# California Department of Transportation : 443
# California Department of Forestry and Fire Protection: 313
# Department of Parks and Recreation: 208
# California Highway Patrol: 107
# California Military Department: 103
```

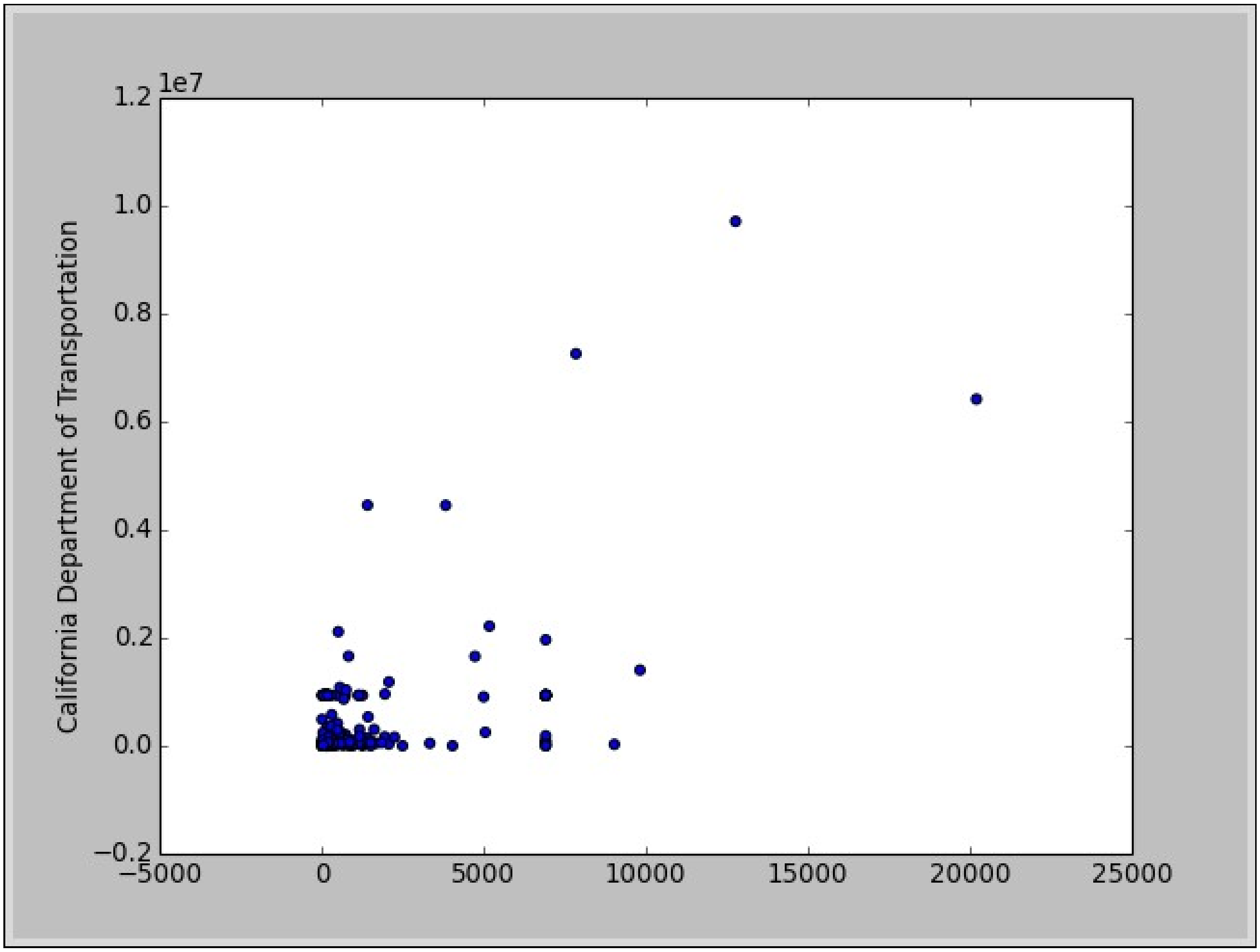
California Department of Forestry and Fire Protection

Pearson Correlation
(0.22651218901438014)

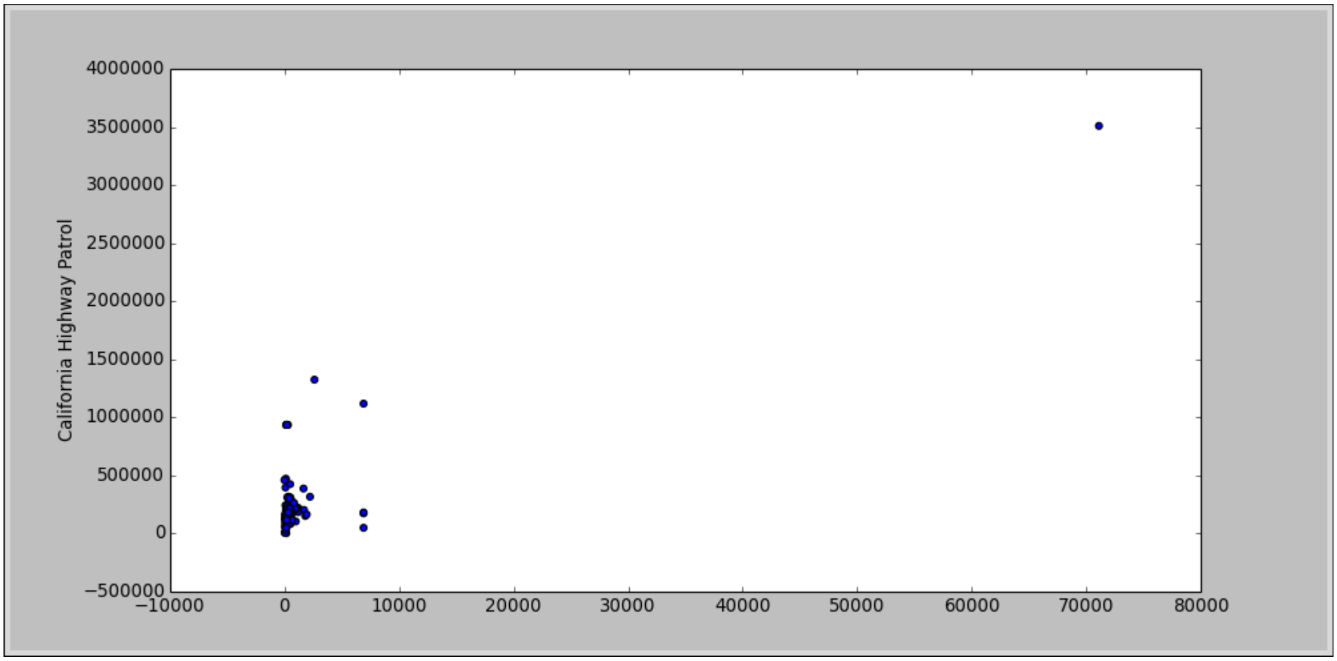


California Department of Transportation

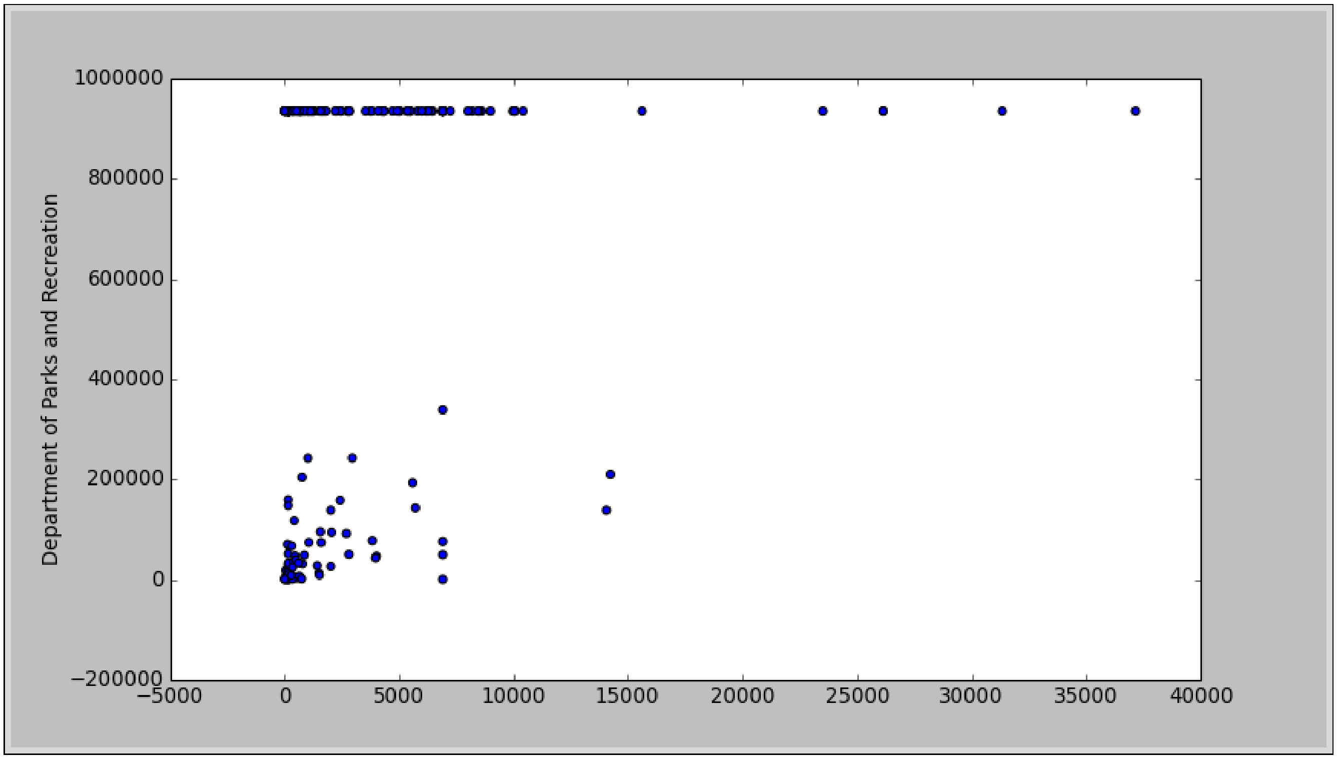
Pearson Correlation
(0.550807088827781)



California Highway Patrol
Pearson Correlation
(0.81679260770973405)

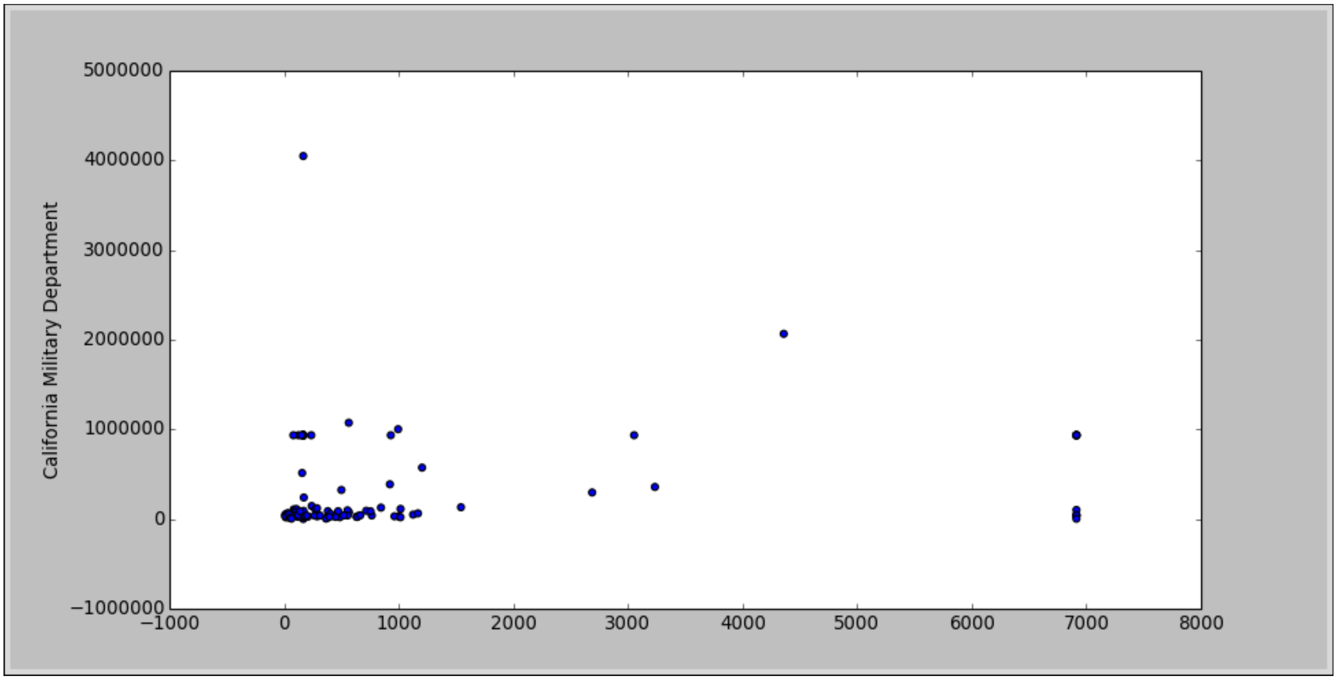


Department of Parks and Recreation:
Pearson Correlation
(0.15526842027669963)



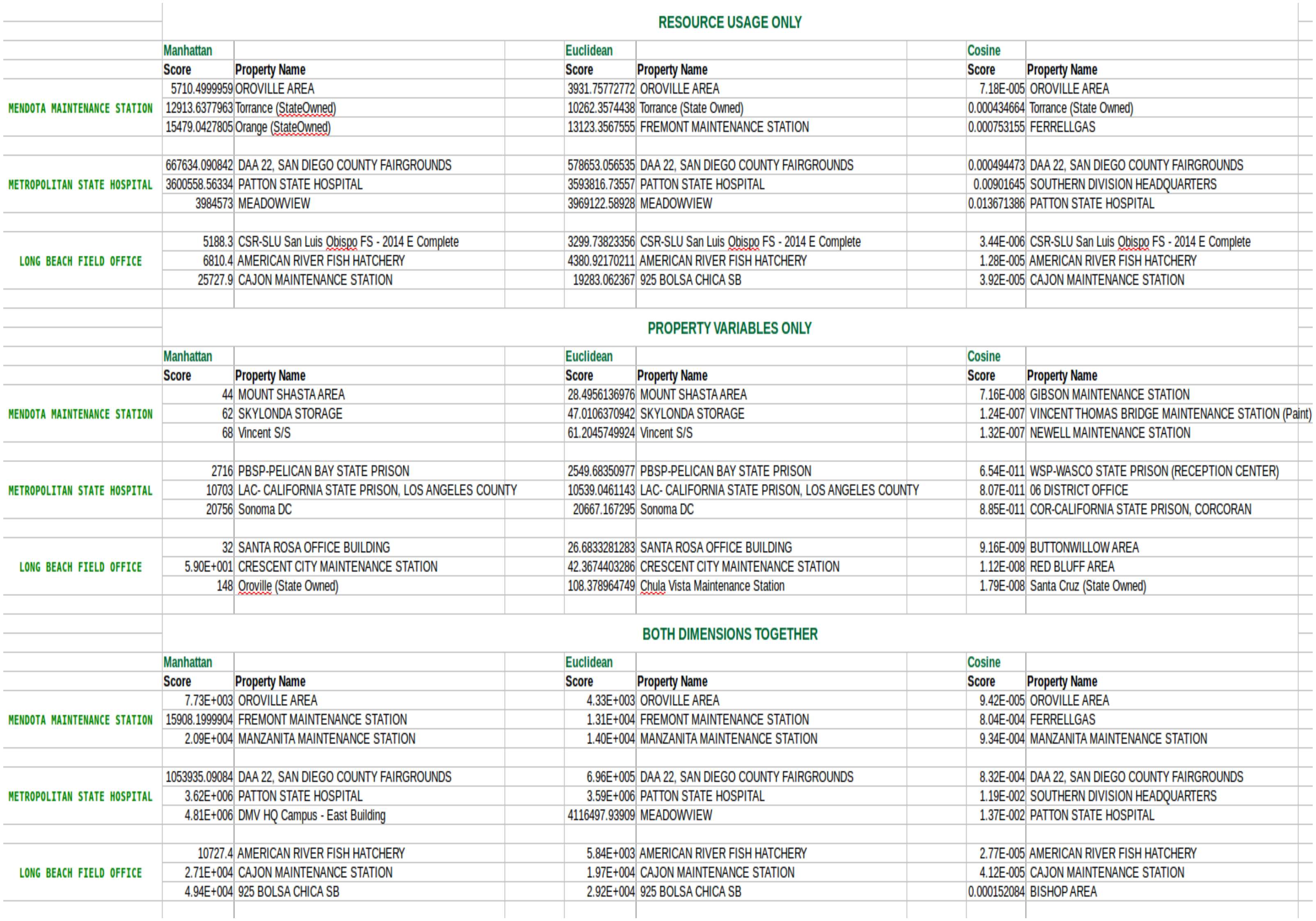
California Military Department

Pearson Correlation
(0.18718108469273598)



Conclusion:
Buildings of *California Department of Transportation* have most correlated Water and Electricity usage and buildings of *Department of Parks and Recreation* have least correlated Water and Electricity usage.

3. Building Similarities.
+ Transformation for nominal data:
- Unique value to each nominal data.



RESOURCE USAGE ONLY

	Manhattan		Euclidean		Cosine	
	Score	Property Name	Score	Property Name	Score	Property Name
MENDOTA MAINTENANCE STATION	5710.4999959	OROVILLE AREA	3931.75772772	OROVILLE AREA	7.18E-005	OROVILLE AREA
	12913.6377963	Torrance (StateOwned)	10262.3574438	Torrance (State Owned)	0.000434664	Torrance (State Owned)
	15479.0427805	Orange (StateOwned)	13123.3567555	FREMONT MAINTENANCE STATION	0.000753155	FERRELLGAS
METROPOLITAN STATE HOSPITAL	667634.090842	DAA 22, SAN DIEGO COUNTY FAIRGROUNDS	578653.056535	DAA 22, SAN DIEGO COUNTY FAIRGROUNDS	0.000494473	DAA 22, SAN DIEGO COUNTY FAIRGROUNDS
	3600558.56334	PATTON STATE HOSPITAL	3593816.73557	PATTON STATE HOSPITAL	0.00901645	SOUTHERN DIVISION HEADQUARTERS
	3984573	MEADOWVIEW	3969122.58928	MEADOWVIEW	0.013671386	PATTON STATE HOSPITAL
LONG BEACH FIELD OFFICE	5188.3	CSR-SLU San Luis Obispo FS - 2014 E Complete	3299.73823356	CSR-SLU San Luis Obispo FS - 2014 E Complete	3.44E-006	CSR-SLU San Luis Obispo FS - 2014 E Complete
	6810.4	AMERICAN RIVER FISH HATCHERY	4380.92170211	AMERICAN RIVER FISH HATCHERY	1.28E-005	AMERICAN RIVER FISH HATCHERY
	25727.9	CAJON MAINTENANCE STATION	19283.062367	925 BOLSA CHICA SB	3.92E-005	CAJON MAINTENANCE STATION

PROPERTY VARIABLES ONLY

	Manhattan		Euclidean		Cosine	
	Score	Property Name	Score	Property Name	Score	Property Name
MENDOTA MAINTENANCE STATION	44	MOUNT SHASTA AREA	28.4956136976	MOUNT SHASTA AREA	7.16E-008	GIBSON MAINTENANCE STATION
	62	SKYLONDA STORAGE	47.0106370942	SKYLONDA STORAGE	1.24E-007	VINCENT THOMAS BRIDGE MAINTENANCE STATION (Paint)
	68	Vincent S/S	61.2045749924	Vincent S/S	1.32E-007	NEWELL MAINTENANCE STATION
METROPOLITAN STATE HOSPITAL	2716	PBSP-PELICAN BAY STATE PRISON	2549.68350977	PBSP-PELICAN BAY STATE PRISON	6.54E-011	WSP-WASCO STATE PRISON (RECEPTION CENTER)
	10703	LAC- CALIFORNIA STATE PRISON, LOS ANGELES COUNTY	10539.0461143	LAC- CALIFORNIA STATE PRISON, LOS ANGELES COUNTY	8.07E-011	06 DISTRICT OFFICE
	20756	Sonoma DC	20667.167295	Sonoma DC	8.85E-011	COR-CALIFORNIA STATE PRISON, CORCORAN
LONG BEACH FIELD OFFICE	32	SANTA ROSA OFFICE BUILDING	26.6833281283	SANTA ROSA OFFICE BUILDING	9.16E-009	BUTTONWILLOW AREA
	5.90E+001	CRESCENT CITY MAINTENANCE STATION	42.3674403286	CRESCENT CITY MAINTENANCE STATION	1.12E-008	RED BLUFF AREA
	148	Oroville (State Owned)	108.378964749	Chula Vista Maintenance Station	1.79E-008	Santa Cruz (State Owned)

BOTH DIMENSIONS TOGETHER

	Manhattan		Euclidean		Cosine	
	Score	Property Name	Score	Property Name	Score	Property Name
MENDOTA MAINTENANCE STATION	7.73E+003	OROVILLE AREA	4.33E+003	OROVILLE AREA	9.42E-005	OROVILLE AREA
	15908.1999904	FREMONT MAINTENANCE STATION	1.31E+004	FREMONT MAINTENANCE STATION	8.04E-004	FERRELLGAS
	2.09E+004	MANZANITA MAINTENANCE STATION	1.40E+004	MANZANITA MAINTENANCE STATION	9.34E-004	MANZANITA MAINTENANCE STATION
METROPOLITAN STATE HOSPITAL	1053935.09084	DAA 22, SAN DIEGO COUNTY FAIRGROUNDS	6.96E+005	DAA 22, SAN DIEGO COUNTY FAIRGROUNDS	8.32E-004	DAA 22, SAN DIEGO COUNTY FAIRGROUNDS
	3.62E+006	PATTON STATE HOSPITAL	3.59E+006	PATTON STATE HOSPITAL	1.19E-002	SOUTHERN DIVISION HEADQUARTERS
	4.81E+006	DMV HQ Campus - East Building	4116497.93909	MEADOWVIEW	1.37E-002	PATTON STATE HOSPITAL
LONG BEACH FIELD OFFICE	10727.4	AMERICAN RIVER FISH HATCHERY	5.84E+003	AMERICAN RIVER FISH HATCHERY	2.77E-005	AMERICAN RIVER FISH HATCHERY
	2.71E+004	CAJON MAINTENANCE STATION	1.97E+004	CAJON MAINTENANCE STATION	4.12E-005	CAJON MAINTENANCE STATION
	4.94E+004	925 BOLSA CHICA SB	2.92E+004	925 BOLSA CHICA SB	0.000152084	BISHOP AREA

```
#dept_names  ##37
#city_names  ##847
#prop_type_names  ##1722
#prop_area_names  ##1557
```

There is no ground truth to measure performance. In this case where data is quantitative, either of the distance measure will perform better. As in the results is shows they have same result for *ALMOST* all the cases.

In terms of dimension, more is better but again there is no ground truth. In this case, I would assume merging both dimension together will perform best.