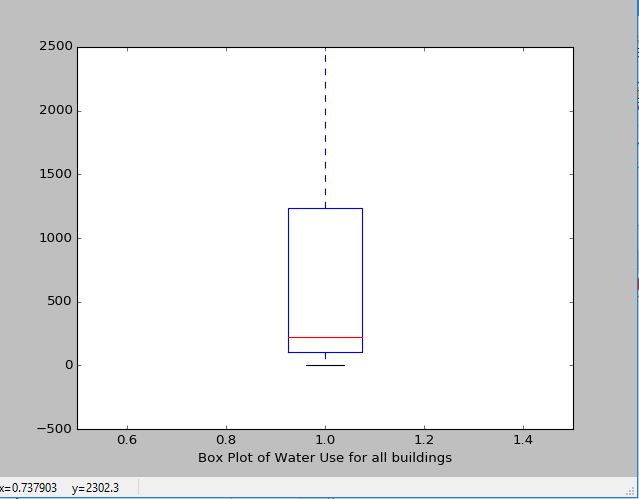
Initially, I have transformed the given raw csv data into excel so that it would become easy for further analysis there by achieving the primary goal of Data Science. There are few missing values (left blank) in the given raw data which I have filled with the mean value of the respective columns. I have tried to substitute the missing values with median, mode and 0s but on careful analysis I found the mean worked well of all those listed above. In other words, I could hardly see any change in the data when I used mean to substitute in the missing values which was not the case with median, mode and 0s.

**Q1. a)** Firstly, I calculated the water usage of all the departments using the central tendencies (Mean, Median and Mode) and using these I plotted the box-plot using the “matplotlib.pyplot” library. Here are the values:

Mean- 6913.163341

Median-225.000000

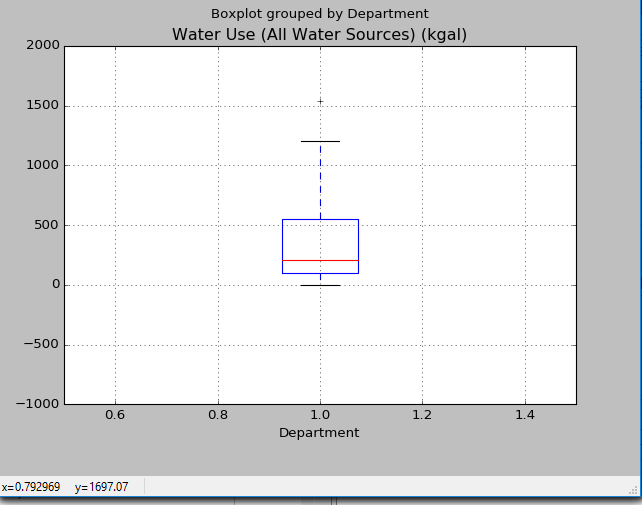
Mode-165.000000



**Fig: Box-plot for all the departments**

Then, I calculated the number of buildings for each department and picked up the top five departments based on the count. Later, using the data and central tendencies (Mean, Median and Mode) of water usage I plotted the box-plot for each of the top five departments in python pandas using the same technique mentioned above. Here are the values and box-plots:

**NOTE:** I have arranged the top five departments in the increasing order of their priority i.e starting with the least department of the top five (CALIFORNIA MILITARY DEPARTMENT), eventually ending with the top department (CALIFORNIA DEPARTMENT OF TRANSPORTATION).

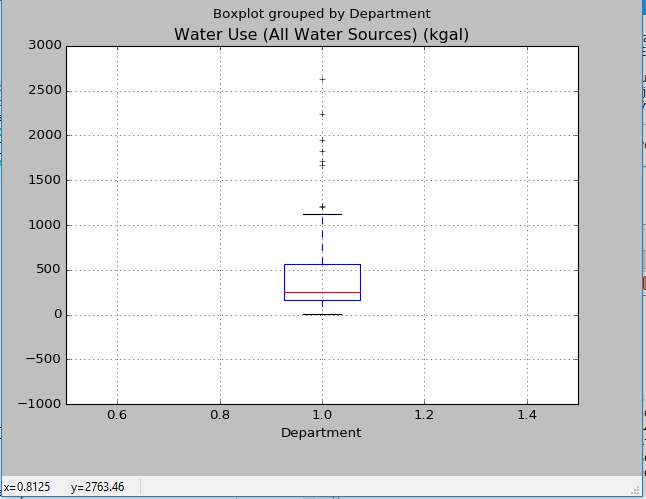


**CALIFORNIA MILITARY DEPARTMENT**

Mean: 772.7516185

Median: 206.4

Mode: 165

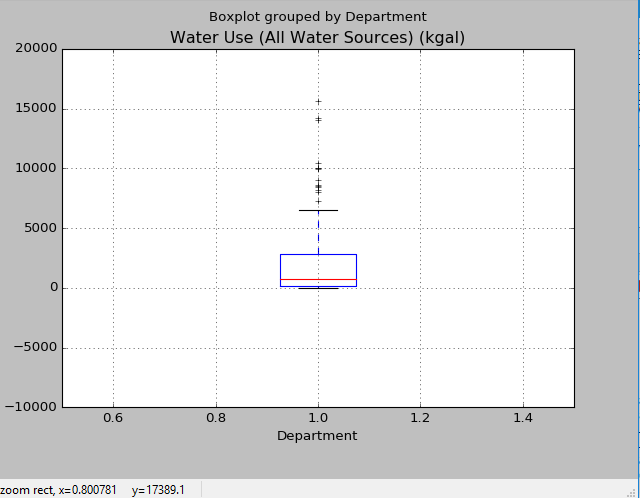


**CALIFORNIA HIGHWAY PATROL**

Mean: 1568.652835

Median: 256.7

Mode: 165

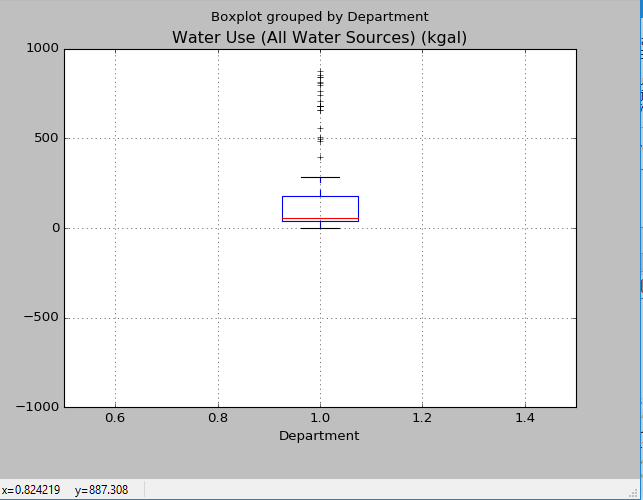


**DEPARTMENT OF PARKS AND RECREATION**

Mean: 2942.108943

Median: 762.65

Mode: 165

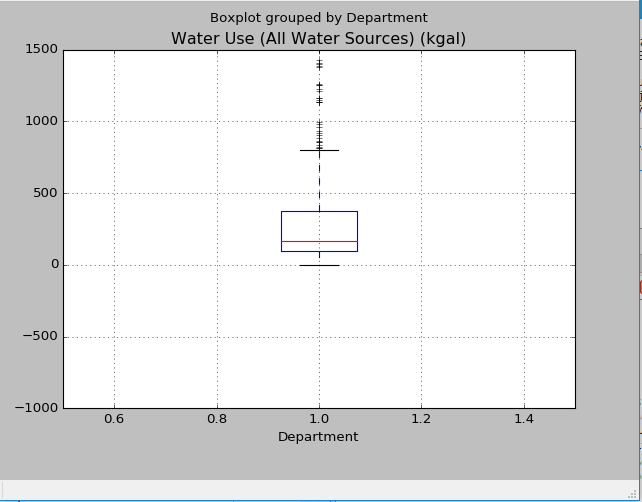


**CALIFORNIA DEPARTMENT OF FORESTRY AND FIRE PROTECTION**

Mean: 785.1817891

Median: 58.6

Mode: 51.5



**CALIFORNIA DEPARTMENT OF TRANSPORTATION**

Mean: 690.5321071

Median:165

Mode: 165

**Q1. b)** After removing outliers taking into account **9 Standard Deviations** from the mean, I obtained the Outlier mean, median and mode values for all the departments and here they are:

**After removing outliers, the central tendencies of all the Departments:**

Outlier Mean- 6027.931806

Outlier Median- 224.600000

Outlier Mode- 165.000000

**After removing outliers, the central tendencies of each of the top five Departments:**

* ***CALIFORNIA DEPARTMENT OF TRANSPORTATION:***

Outlier Mean- 457.833662

Outlier median is 165.000000

Outlier mode is 165.000000

* ***CALIFORNIA DEPARTMENT OF FORESTRY AND FIRE PROTECTION:***

Outlier mean is 454.710897

Outlier median is 58.600000

Outlier mode is 51.500000

* ***DEPARTMENT OF PARKS AND RECREATION***

Outlier mean is 2698.983163

Outlier median is 762.650000

Outlier mode is 165.000000

* ***CALIFORNIA HIGHWAY PATROL***

Outlier mean is 467.709498

Outlier median is 256.250000

Outlier mode is 165.000000

* ***CALIFORNIA MILITARY DEPARTMENT***

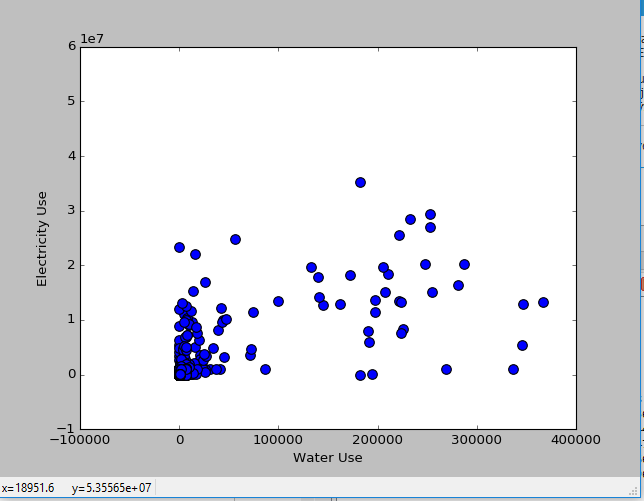
Outlier mean is 459.465306

Outlier median is 206.400000

Outlier mode is 165.000000

**OBSERVATION:** On comparing the central tendency values before and after removing the outliers, I found that the mean values have decreased, median values remained almost the same and the mode values are exactly the same. This made the data more standardized.

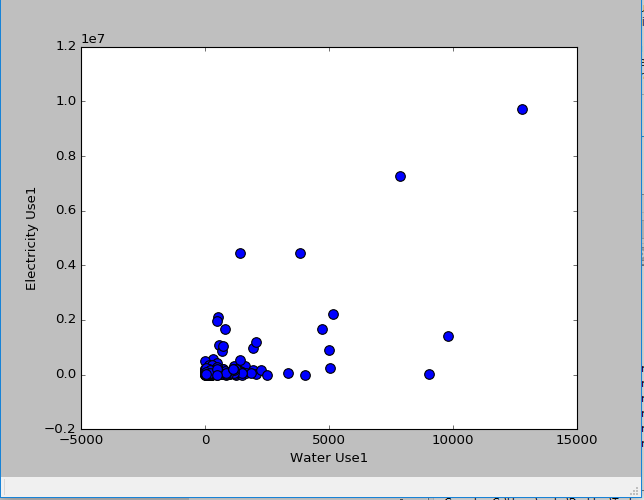
**Q2. a)** Considering the “Water Use” on X-Axis and “Electricity” on Y-Axis, I drew a scatter plot for all the departments and here is the figure:



**Fig: Scatter plot for all the departments**

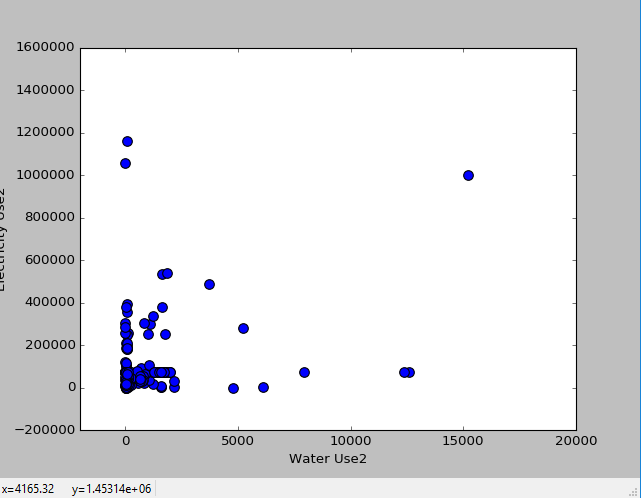
**Pearson Correlation Coefficient is**: 0.666118

Considering the “Water Use” on X-Axis and “Electricity” on Y-Axis, I drew a scatter plot for each of the top five departments that I have listed above. Here are the figures:



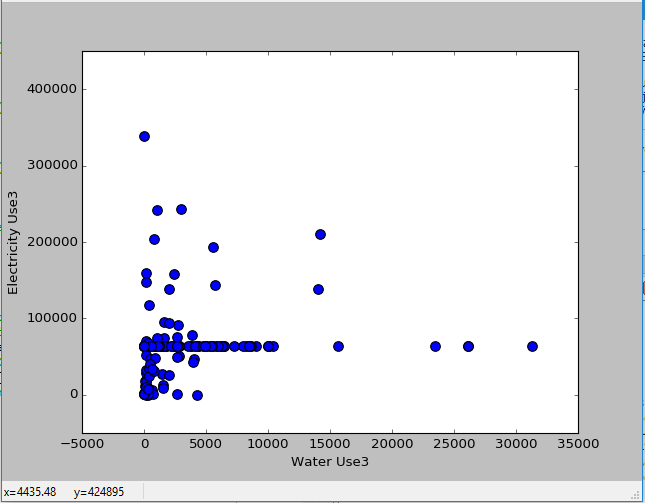
**Fig 1: Scatter plot for California Department of Transportation**

**Pearson Correlation Coefficient for this department is:** 0.731836



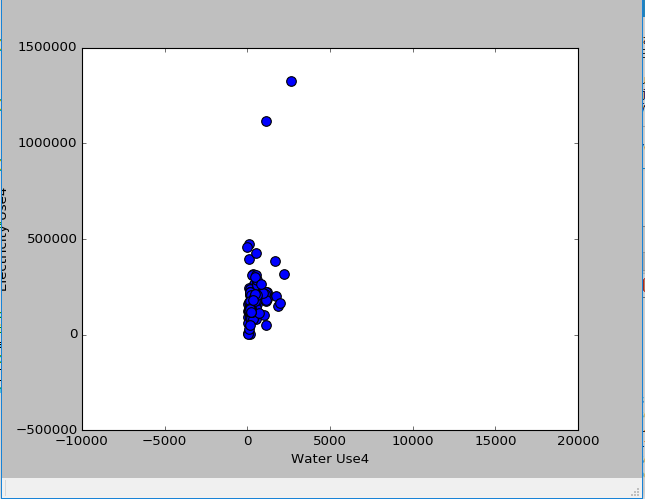
**Fig 2: California Department of Forestry and Fire Protection**

**Pearson Correlation Coefficient for this department is:** 0.299044



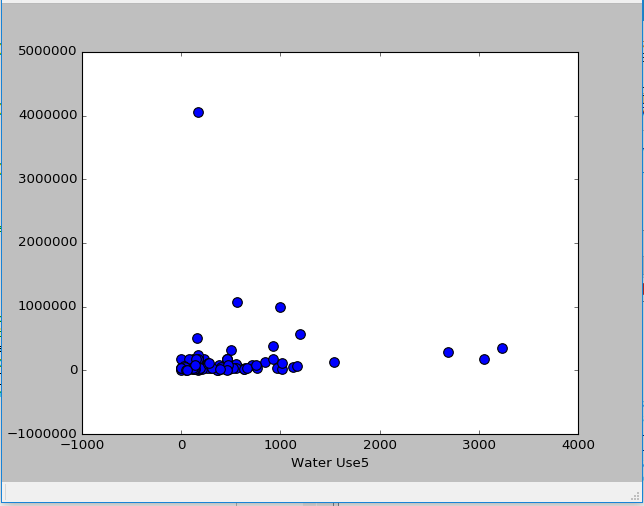
**Fig 3: Department of Parks and Recreation**

**Pearson Correlation Coefficient for this department is:** 0.108347



**Fig 4: California Highway Patrol**

**Pearson Correlation Coefficient of this department is:** 0.897666



**Fig 5: California Military Department**

**Pearson Correlation Coefficient of this department is:** 0.286496

**NOTE**:

On observation, I found that the water usage and electricity are **Positively and highly co-related** for **California Department of Transportation** and **California Highway Patrol** as the values are **closer to ONE (1).** For the other three departments namely **California Military Department, Department of Parks and Recreation and California Department of Forestry and Fire Protection**,the water usage and electricity are **Positively and weakly co-related** as the values are **closer to ZERO (0).**

**Q3)** Using two distance metrics (Euclidean and Manhattan) and one similarity function (Cosine), find the three buildings similar to following buildings.

**Dimension 1:**

***Euclidean* for MENDOTA MAINTENANCE STATION MAINTENANCE STATION**

***Distance Property Name***

3931.7577277243704 OROVILLE AREA

10262.35744383182 Torrance (State Owned)

13123.35675546162 FREMONT MAINTENANCE STATION

***Cosine Similarity* for MENDOTA MAINTENANCE STATION MAINTENANCE STATION**

***Distance Property Name***

0.9999282310451871 OROVILLE AREA

0.9995653361810225 Torrance (State Owned)

0.9991610397447368 Orange (State Owned)

***Manhattan* for MENDOTA MAINTENANCE STATION MAINTENANCE STATION**

***Distance Property Name***

5710.4999959000015 OROVILLE AREA

12913.63779629998 Torrance (State Owned)

15479.042780500007 Orange (State Owned)

***Euclidean* for METROPOLITAN STATE HOSPITAL STATE HOSPITAL**

***Distance Property Name***

578653.0492122686 DAA 22, SAN DIEGO COUNTY FAIRGROUNDS

3593816.735562838 PATTON STATE HOSPITAL

3969122.5892776507 MEADOWVIEW

***Cosine Similarity* for** **METROPOLITAN STATE HOSPITAL STATE HOSPITAL**

***Distance Property Name***

0.9995055267155946 DAA 22, SAN DIEGO COUNTY FAIRGROUNDS

0.9863286134013611 PATTON STATE HOSPITAL

0.9828436469705564 DMV HQ Campus - East Building

***Manhattan*** **for** **METROPOLITAN STATE HOSPITAL STATE HOSPITAL**

***Distance Property Name***

667634.0055999989 DAA 22, SAN DIEGO COUNTY FAIRGROUNDS

3600558.5599999987 PATTON STATE HOSPITAL

3984573.0000000005 MEADOWVIEW

***Euclidean* for LONG BEACH FIELD OFFICE**

***Distance Property Name***

3299.7382335573016 CSR-SLU San Luis Obispo FS - 2014 E Complete

4380.921702107876 AMERICAN RIVER FISH HATCHERY

19283.062367010112 925 BOLSA CHICA SB

***Cosine* for LONG BEACH FIELD OFFICE**

***Distance Property Name***

0.9999965570414939 CSR-SLU San Luis Obispo FS - 2014 E Complete

0.9999872075862348 AMERICAN RIVER FISH HATCHERY

0.9999607755106035 CAJON MAINTENANCE STATION

***Manhattan* for LONG BEACH FIELD OFFICE**

***Distance Property Name***

5188.299999999971 CSR-SLU San Luis Obispo FS - 2014 E Complete

6810.39999999996 AMERICAN RIVER FISH HATCHERY

25727.899999999896 CAJON MAINTENANCE STATION

**DIMENSION 2:**

***Euclidean* for MENDOTA MAINTENANCE STATION**

***Distance Property Name***

137.16413525408163 1317 15TH ST LOMBARD APARTMENT COMPLEX

141.04609175726918 NACYF-N.A. CHADERJIAN YOUTH CORRECTIONAL FACILITY(NCYCC)

179.9638852659055 El Centro (State Owned)

***Cosine Similarity* for** **MENDOTA MAINTENANCE STATION**

***Distance Property Name***

0.9999999920225857 OROVILLE WA

0.9999999370867985 011 Bateson Building

0.999999929790026 CUYAMA MAINTENANCE STATION

***Manhattan* for** **MENDOTA MAINTENANCE STATION**

***Distance Property Name***

196.0 1317 15TH ST LOMBARD APARTMENT COMPLEX

202.0 NACYF-N.A. CHADERJIAN YOUTH CORRECTIONAL FACILITY(NCYCC)

227.0 011 Bateson Building

***Euclidean* for METROPOLITAN STATE HOSPITAL**

***Distance Property Name***

77.62087348130012 CSR-TCU Columbia HB (AAB) 2014 E MPP Local govt

80.21845174272562 PINNACLE PETROLEUM

137.20422734012243 Bolsa Chica

***Cosine Similarity* for** **METROPOLITAN STATE HOSPITAL**

***Distance Property Name***

0.9999998971260169 SACRAMENTO LA MANCHA FIELD OFFICE

0.9999998762152188 OCEANSIDE OFFICE BUILDING

0.9999998401031773 WESTMINSTER OFFICE BUILDING

***Manhattan*** **for** **METROPOLITAN STATE HOSPITAL**

***Distance Property Name***

117.0 PINNACLE PETROLEUM

133.0 CSR-TCU Columbia HB (AAB) 2014 E MPP Local govt

175.0 NEVADA CITY (SUTTER/SIERRA REGION) MAINTENANCE

***Euclidean* for LONG BEACH FIELD OFFICE**

***Distance Property Name***

38.768543949960254 NAPA FIELD OFFICE

115.04781614615725 Willow Springs Maint

115.58979193683152 CTC-RICHARD A. MCGEE CORRECTIONAL TRAINING CENTER

***Cosine Similarity* for** **LONG BEACH FIELD OFFICE**

***Distance Property Name***

0.9999998062067453 Vidal Maint

0.9999994696136498 VALLEJO MAINTENANCE STATION

0.9999993625856142 Fresno Maintenance Station

***Manhattan* for** **LONG BEACH FIELD OFFICE**

***Distance Property Name***

67.0 NAPA FIELD OFFICE

151.0 CTC-RICHARD A. MCGEE CORRECTIONAL TRAINING CENTER

164.0 TEHACHAPI SAND STORAGE

**DIMENSION 3:**

***Euclidean* for MENDOTA MAINTENANCE STATION**

***Distance Property Name***

20083.45066032518 MANZANITA MAINTENANCE STATION

20211.13841231573 ONTARIO MAINTENANCE STATION

21878.45543859249 CNR-LNU Cloverdale FS - 2014 E N Complete

***Cosine Similarity* for** **MENDOTA MAINTENANCE STATION**

***Distance Property Name***

0.9978839464540826 CNR-LNU Cloverdale FS - 2014 E N Complete

0.9976667936610482 MANZANITA MAINTENANCE STATION

0.9975079889292003 ONTARIO MAINTENANCE STATION

***Manhattan* for** **MENDOTA MAINTENANCE STATION**

***Distance Property Name***

28155.49959 OROVILLE AREA

32567.645808200003 MANZANITA MAINTENANCE STATION

34676.94968730001 1228 O STREET GIBSON ARMS

***Euclidean* for METROPOLITAN STATE HOSPITAL**

***Distance Property Name***

578723.9105952475 DAA 22, SAN DIEGO COUNTY FAIRGROUNDS

3594002.6257510628 PATTON STATE HOSPITAL

3969123.0331840464 MEADOWVIEW

***Cosine Similarity* for** **METROPOLITAN STATE HOSPITAL**

***Distance Property Name***

0.9995052975236706 DAA 22, SAN DIEGO COUNTY FAIRGROUNDS

0.9863260440828219 PATTON STATE HOSPITAL

0.9828332081125152 DMV HQ Campus - East Building

***Manhattan* for** **METROPOLITAN STATE HOSPITAL**

***Distance Property Name***

676884.0055999989 DAA 22, SAN DIEGO COUNTY FAIRGROUNDS

3637438.5599999987 PATTON STATE HOSPITAL

3986681.0000000005 MEADOWVIEW

***Euclidean* for LONG BEACH FIELD OFFICE**

***Distance Property Name***

7238.360930486941 AMERICAN RIVER FISH HATCHERY

9947.114376038906 CSR-SLU San Luis Obispo FS - 2014 E Complete

19838.675386476698 CAJON MAINTENANCE STATION

***Cosine Similarity* for** **LONG BEACH FIELD OFFICE**

***Distance Property Name***

0.9999536849209146 AMERICAN RIVER FISH HATCHERY

0.9999501487703771 CAJON MAINTENANCE STATION

0.9999071755083188 CSR-SLU San Luis Obispo FS - 2014 E Complete

***Manhattan* for** **LONG BEACH FIELD OFFICE**

***Distance Property Name***

12920.39999999996 AMERICAN RIVER FISH HATCHERY

14818.29999999997 CSR-SLU San Luis Obispo FS - 2014 E Complete

28922.899999999896 CAJON MAINTENANCE STATION