DS 3001
Homework 1

Carlos Barcelos
March 25, 2018

Task 1

This task required the analysis of **water usage** and its usage across **departments**. For this reason, I preprocessed the original data file to only include information about these two key attributes. The dataset contained the following attributes: *Department* and *Water Use*. This was accomplished by reading in the original dataset and dropping all columns besides for those three listed above.

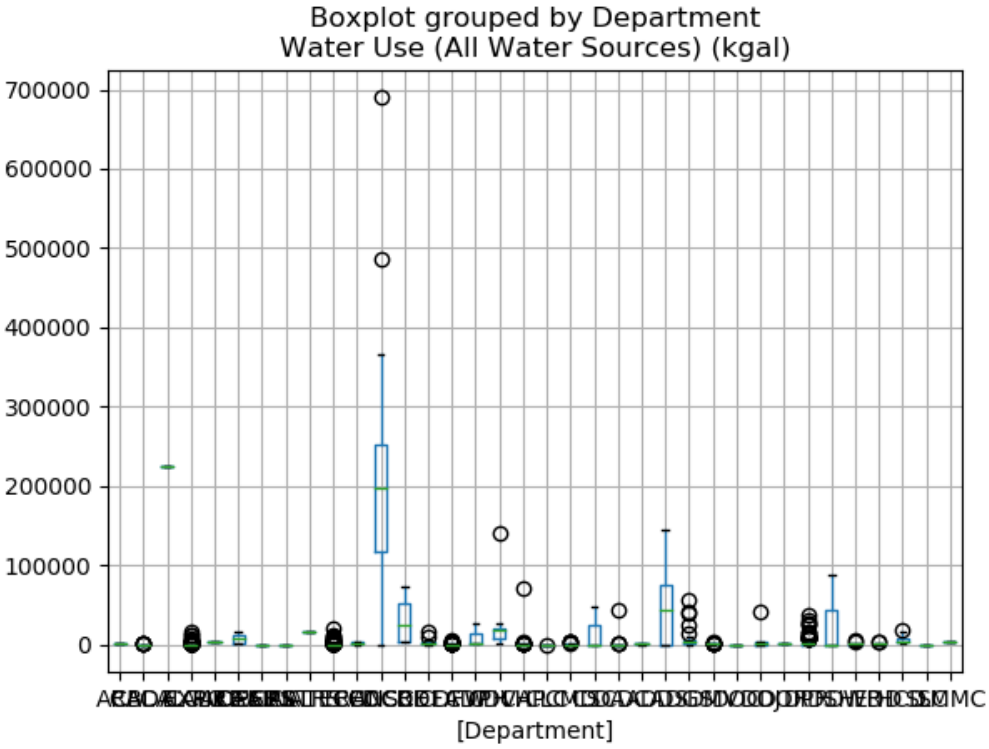
Next, missing values were removed. Because this task involves only one attribute, it was an easy decision to simply remove rows with missing values as they would not contribute to the overall results. This was accomplished with pandas with a simple for loop which removed all records with missing data values (NaN) in their Water Use column. With this clean data frame, it was possible to compute measures of central tendency and create box-plots.

Central Tendency and Box Plots for all buildings

Mean (kgal) = 6913.163341

Median (kgal) = 182.35

Mode (kgal) = 165.0

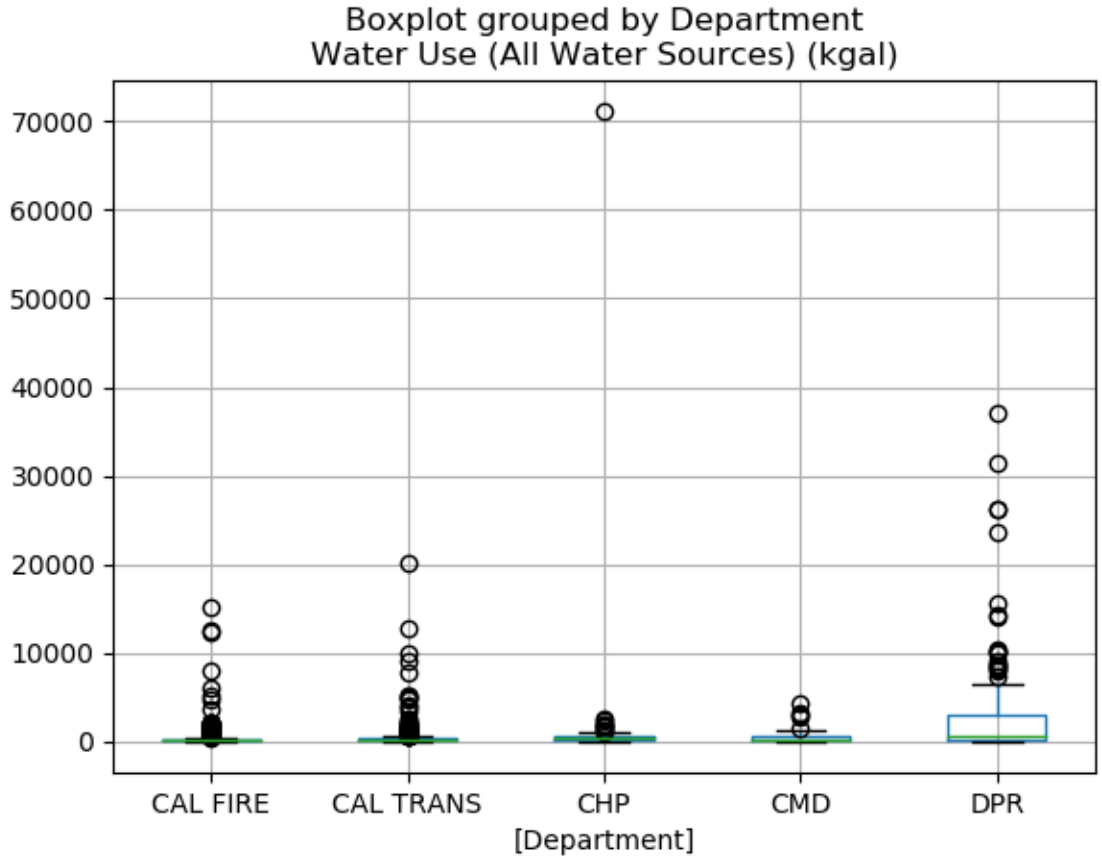


Central Tendency and Box Plots for top 5 departments by usage

The next step required finding the top five departments by number of buildings associated with them. To find these departments, I first needed to take the original data frame and include only the top five departments by building count. To do this, I first counted the total number of buildings associated with each department. Then I could sort those and remove all other departments that were not in the top five departments by building count. Finally, the same mean, median, mode, and box plot computations could be applied to these top five departments:

Rank	*Department*	*Mean (kgal)*	*Median (kgal)*	*Mode (kgal)*
1	CAL TRANS	502.406047	165.00	165.0
2	CAL FIRE	501.881789	58.60	51.5
3	DPR	2698.983163	672.85	165.0
4	CHP	1127.776699	255.30	165.0
5	CMD	459.465306	172.05	165.0

Table 1 Central tendency measures of top five departments



Central Tendency and Box Plots for top 5 departments by usage (without outliers)

Outliers were removed for being too extreme and skewing the data. Outliers were defined as being a value higher or lower than 1.5 times the Inter-quartile range (IQR). The IQR was calculated as IQR = Q3 – Q1. The values of Q1 and Q3 could easily be calculated using Pandas `quantile()` function:

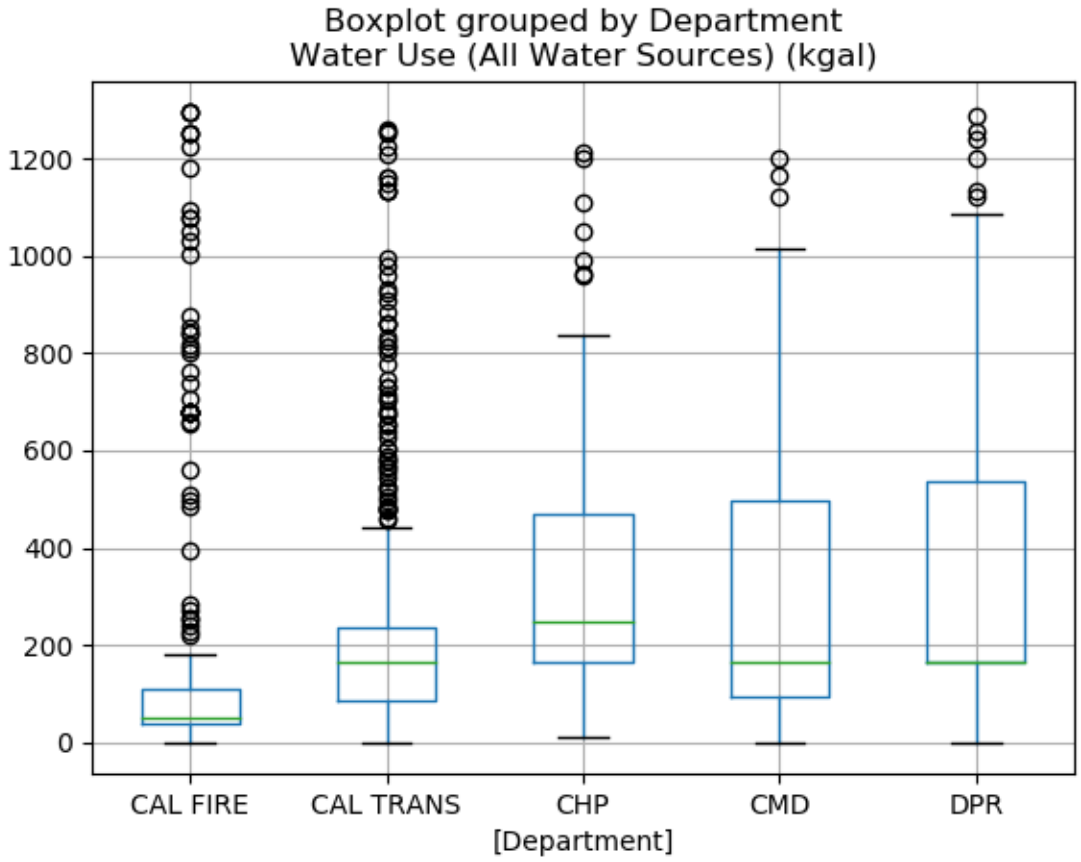
$$Outlier > Q3 + 1.5 * IQR$$

OR

$$Outlier < Q1 - 1.5 * IQR$$

Rank	*Department*	*Mean (kgal)*	*Median (kgal)*	*Mode (kgal)*
1	CAL TRANS	238.988030	165.0	165.0
2	CAL FIRE	174.770213	51.5	51.5
3	DPR	353.985246	165.0	165.0
4	CHP	344.269792	245.9	165.0
5	CMD	324.241935	165.0	165.0

Table 2 Central tendency measures of top five departments (no outliers)



Without the outliers, the mean water usage for all departments is significantly smaller. This implies that all departments had at some outliers in their data which drove up their average water usage. The median water usage seems similar for all departments besides for DPR which now has a much smaller mean. This implies that the DPR department had data heavily skewed towards the high end of water usage. The mode of each department is unchanged which is expected; none of the outliers were the most frequent data point or else they would not have been outliers.

Task 2

This task required the analysis of **water usage** and its usage across **departments**. For this reason, I preprocessed the original data file to only include information about these two key attributes. The dataset contained the following attributes: *Department* and *Water Use*. This was accomplished by reading in the original dataset and dropping all columns besides for those three listed above.

Next, missing values were removed. Because this task involves only one attribute, it was an easy decision to simply remove rows with missing values as they would not contribute to the overall results. This was accomplished with pandas with a simple for loop which removed all records with missing data values (NaN) in their Water Use column. With this clean data frame, it was possible to compute measures of central tendency and create box-plots.

This task required comparing the relationship between a building's **water usage** and its **electricity use**. It then required a similar water usage vs. electricity use comparison to be run across the top five **departments**. For this reason, I preprocessed the original data file to only include information about these four key attributes. The dataset contained the following attributes: *Department, Water Use*, and *Electricity Use*. This was accomplished by reading in the original dataset and dropping all columns besides for those three listed above. The final step of preprocessing involved dropping all empty rows like Task 1.

The relationship between Water Use and Electricity Use of a building is displayed below. The Person's correlation coefficient of this relationship is 0.700323. This implies that there is a moderately strong positive correlation between these two values

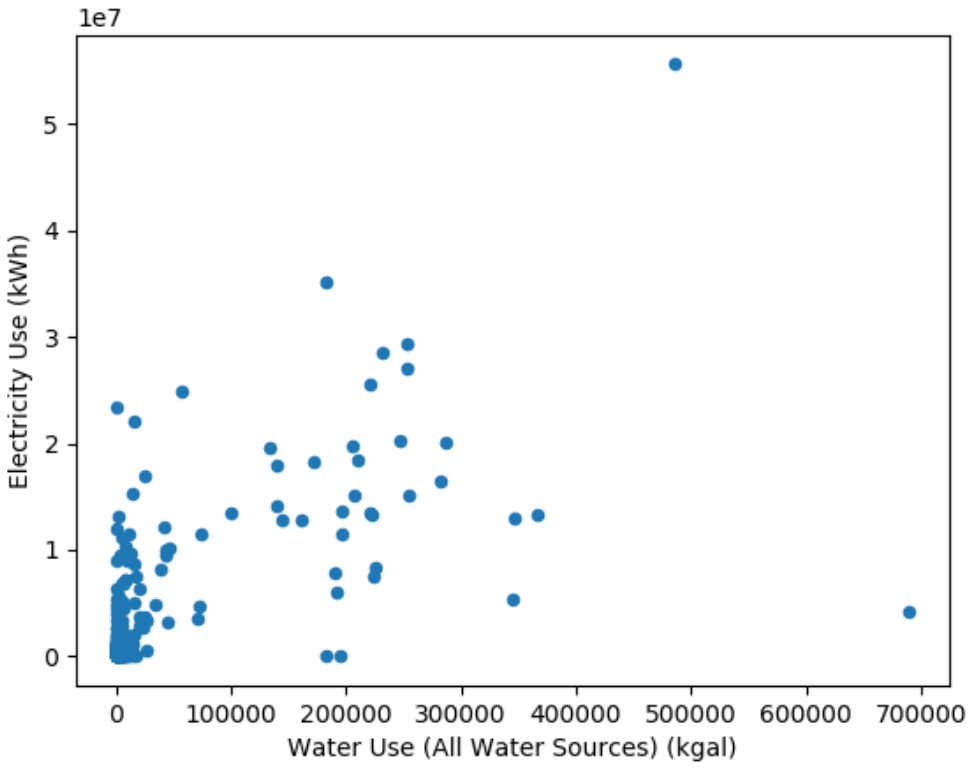


Figure 1 Water v. Electricity (All Departments)

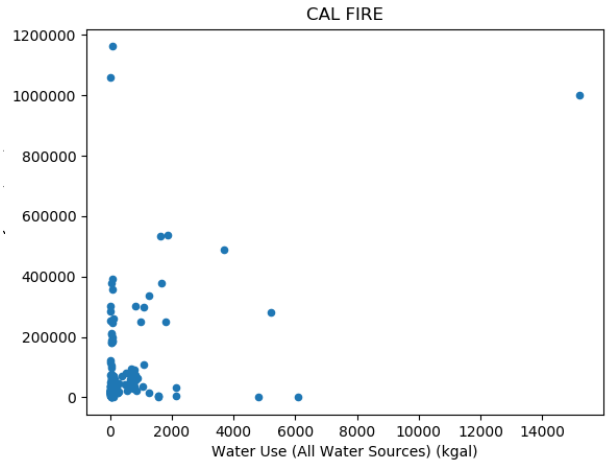
Correlation for the top five departments

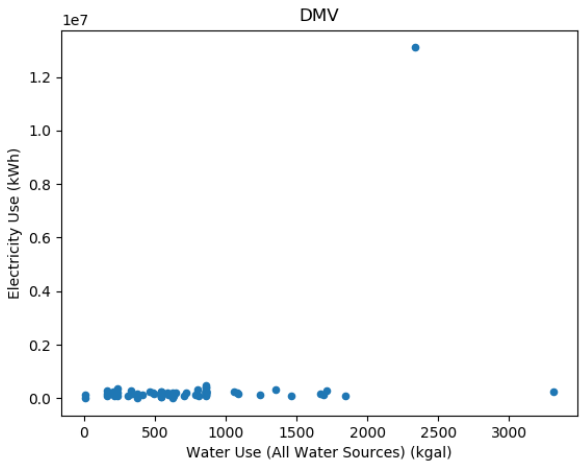
The top five departments were found in a similar fashion to Task 1. The Person's correlation coefficients of each of the top five departments and their respective scatter plots are below.

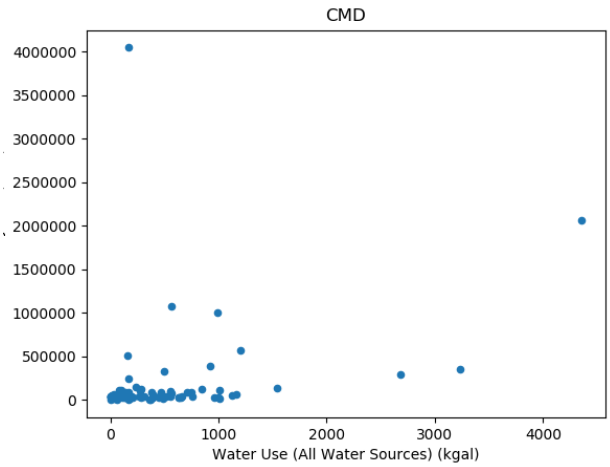
Based on the scatterplots and correlation coefficients it is clear that some department's correlation between water and electricity use is stronger than others. For example, the CHP department has the strongest correlation of the group at 0.92 while the CMD has the weakest correlation at 0.31. With this information it is possible to conclude that one must be careful concluding that in general water usage and electrical usage are highly correlated. As we can see, with the various correlation metrics, that statement is not always true. While general trends are important, one must be diligent and consider the specifics of the data to be accurate.

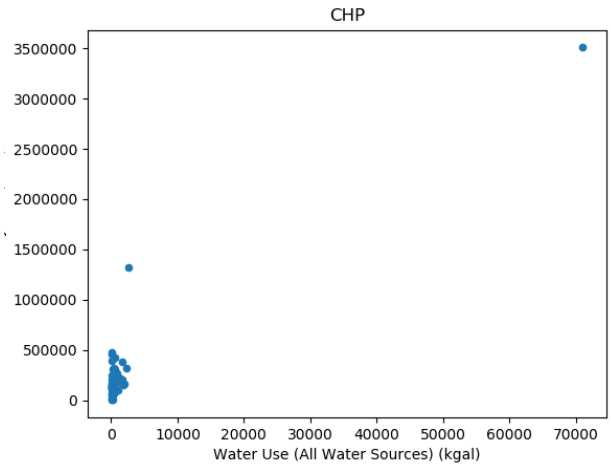
Rank	*Department*	*Person's correlation coefficient*
1	CAL TRANS	0.740944
2	CAL FIRE	0.432206
3	DMV	0.363594
4	CHP	0.926501
5	CMD	0.315039

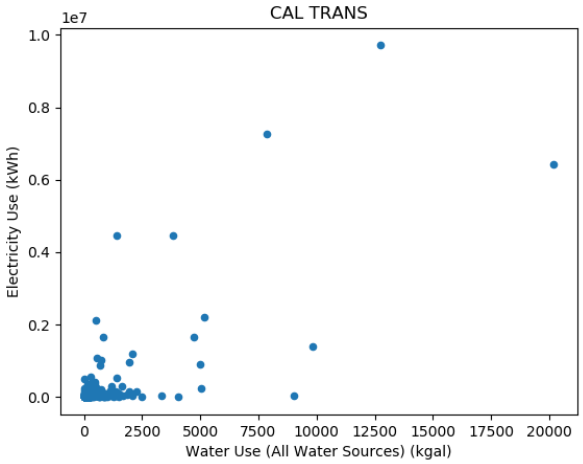
Table 3 Correlation of water usage v. electricity usage of top five departments


CAL FIRE
1200000
1000000
800000
600000
400000
200000
0
0
2000
4000
6000
8000
10000
12000
14000
Water Use (All Water Sources) (kgal)


DMV
1e7
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Electricity Use (kWh)
0
500
1000
1500
2000
2500
3000
Water Use (All Water Sources) (kgal)


CMD
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0
1000
2000
3000
4000
Water Use (All Water Sources) (kgal)


CHP
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0
10000
20000
30000
40000
50000
60000
70000
Water Use (All Water Sources) (kgal)


CAL TRANS
1e7
1.0
0.8
0.6
0.4
0.2
0.0
Electricity Use (kWh)
0
2500
5000
7500
10000
12500
15000
17500
20000
Water Use (All Water Sources) (kgal)

Task 3

This task had two parts. It was completed by reading in the initial CSV file and then accomplishing each part, **Resource Usage Based Similarity** and **Property Variable Based Similarity**, as separate problems. The result of each problem is listed below.

Resource Usage Based Similarity

To find similarity based on resource usage I kept only the following variables in the data frame: Property Name, Electricity Use, Natural Gas Use, Propane Use, Water Use, and Site Energy Use. Property Name was only used for the result indexing while the remaining variables were all quantitative meaning that there was no need to transform them for the following analysis.

The next step was dealing with missing values. The method of dropping rows with missing values was unacceptable due to how many columns and missing values there were. If a row was dropped due to it containing one missing value, the original data frame of 1722 rows would be reduced to only 6 rows. This is a 99.7% reduction in the data set – unacceptable. Instead, I decided to fill all missing values with the average of their column. For example, if a row was missing the Propane Use value it would be replaced with the average of all existing Propane Use values.

By examining the results of the process below, it is clear that the 'OROVILLE AREA' and 'Torrance (State Owned)' are the most similar because they are unanimously listed first and second respectively by all three methods. But the third most similar building is up to debate. All three methods report a unique property for the third most similar building leading me to conclude that we can not draw any certain conclusions from these varying results.

Euclidean Distance

Euclidean distance yields the following as the most similar properties to the search property:

0) MENDOTA MAINTENANCE STATION (perfect correlation because of being the search property)
1) OROVILLE AREA
2) Torrance (State Owned)
3) FREMONT MAINTENANCE STATION

Manhattan Distance

Manhattan distance yields the following as the most similar properties to the search property:

0) MENDOTA MAINTENANCE STATION (perfect correlation because of being the search property)
1) OROVILLE AREA
2) Torrance (State Owned)
3) Orange (State Owned)

Cosine Similarity

Cosine similarity yields the following as the most similar properties to the search property:

0) MENDOTA MAINTENANCE STATION (perfect correlation because of being the search property)
1) OROVILLE AREA
2) Torrance (State Owned)
3) FERRELLGAS

Property Variable Based Similarity

To find similarity based on property details I kept only the following variables in the data frame: Department Name, Property Name, City, Primary Property Type, and Property Area. Property Name was only used for the result indexing while the remaining variables were all nominal meaning that they needed to be transformed to complete this analysis.

Before transforming the nominal data, it was necessary to remove any NaN values. This was accomplished easily by using the build in Pandas `dropna()` function. This reduced the dataset from 1722 to 1395 entries. This was a 18.9% reduction in the data set – acceptable.

The nominal attributes were then transformed by using the get_dummies() function. This function performs one hot encoding by creating a unique value for each answer within an attribute. The sub-values are then encoded with a 0 or 1 (one hot) such that there are as many sub-categories as there are unique responses to the attribute. For example, the following *Fruit* column with responses [apple, orange, banana] would be transformed to three new *isFruit* columns {isApple, isOrange, isBanana}.

By examining the results of the three similarity metrics, it stands out that the cosine similarity results rather different buildings than either the Euclidean or Manhattan distance. In fact, the Euclidean and Manhattan distance measure report three most similar buildings in order leading me to believe that their result is correct while something may have been off while using cosine similarity to calculate the similarity for a record with largely categorical data. For this reason, I am confident in reporting that the 'SANTA BARBARA OFFICE BUILDING' is the most similar building to the original while 'LONG BARN MAINTENANCE STATION' and 'KINGVALE MAINTENANCE STATION' are the second and third most similar building respectively.

Euclidean Distance

Euclidean distance yields the following as the most similar properties to the search property:

0) MENDOTA MAINTENANCE STATION (perfect correlation because of being the search property)
1) SANTA BARBARA OFFICE BUILDING
2) LONG BARN MAINTENANCE STATION
3) KINGVALE MAINTENANCE STATION

Manhattan Distance

Manhattan distance yields the following as the most similar properties to the search property:

0) MENDOTA MAINTENANCE STATION (perfect correlation because of being the search property)
1) SANTA BARBARA OFFICE BUILDING
2) LONG BARN MAINTENANCE STATION
3) KINGVALE MAINTENANCE STATION

Cosine Similarity

Cosine similarity yields the following as the most similar properties to the search property:

0) MENDOTA MAINTENANCE STATION (perfect correlation because of being the search property)
1) California Veterans Home, Ventura
2) RICHMOND LAB. AND OFFICE BLDGS A, B, C, D, E, F, G, H, J, K
3) SKYLONDA STORAGE