Homework #1

DS 3001, D-term 2018

---

100 points total [8% of your final grade]

Due: March 25, 2018 by 11:59pm
[no submission will be accepted after March 28, 2018 at 11:59pm]

Delivery: Submit via canvas

Overview

As the first step on your way to becoming a full-fledged data scientist, you decide to explore a California State Buildings Sustainability dataset:
Link: http://web.cs.wpi.edu/~kmlee/ds3001/CAStateBuildingMetrics.csv

You face three related tasks that will require you to transform the raw data into a more usable form, and then perform a series of analyses, toward better understanding your data.

Tasks

For each task, you should describe how you transformed the raw csv data to the format you can perform your analysis on. Describe the issues you faced (e.g., missing values), how you resolved these issues, and justify why you decided to handle it that way.

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.

- 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 based on the number of buildings associated with). You should plot box-plots for all buildings, as well as for the individual departments.
- 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?

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

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

- 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?

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

Property Name
MENDOTA MAINTENANCE STATION

You should use the following dimensions in the distance measurement:

- Resource Usage only: Electricity Use, Natural Gas Use, Propane Use, Water Use, Site Energy Use
- Property variables only : Department Name,City, Primary Property Type, Property Area

You should be careful in how you measure distance, since some dimensions are nominal, while others are quantitative. You may need to apply (and explain) any transformations you deem necessary to conduct your analyses.

Analyze your results to see which of the (similarity measure, dimension) pair performs the best. Explain your reasoning and what you can conclude.

What to turn in:

- You should turn in ONE file: a PDF report with your answers to the homework questions (hw1_yourname.pdf).
- This is an individual assignment, but you may discuss general strategies and approaches with other members of the class (refer to the syllabus for details of the homework collaboration policy). At the top of your report, please write the names of classmates you consulted and the nature of your discussion.