STATS506 Final Project: Lighting Types

Heleyna Tucker

Introduction

The data used in this research project was from the 2018 Commercial Buildings Energy Consumption Survey. About 6,400 buildings were surveyed. Participants were asked a series of questions that required either numerical grouping (1 = Yes, 2 = No) or a numerical value (usually from 0-100) to answer each question. For my topic of interest, I will focus on percentage data for the following variables: lights on during no hours, lights on when open, lights on during off hours, lit by fluorescent, lit by compact fluorescent (CFL) bulbs, lit by incandescent, lit by halogen, lit by HID, lit by LED, and lit by other lighting. Using this data, the following question will be answered: Is there a difference in percentage of lighting types based on how often lights are kept on during various hours? This question will be answered with K-Means clustering.

Approach

The first steps required before doing any data analysis is data cleaning and examining from the dataset provided. The needed variables were pulled from the original dataset and renamed for ease of understanding. The building IDs and the percent lighting data described above were used to create a new, compressed dataset. It would be expected that every percentage of lighting types for each building would add up to 100% to check this, there was a new variable added to this dataset that totalled the lighting type percentages. It was found that about 80% of the buildings had lighting percentage data that added up to 100. For the sake of simplicity and accuracy, the buildings that had inaccurate percentages were omitted from the compressed dataset. Furthermore, the values that were recorded as NA were replaced with zeros throughout the compressed dataset because these values were either considered missing or not applicable.

A summary of this compressed data was done to get an overview of simple statistics (see Appendix Figure A1). To find that there was not much data for the lights on during no hours category. Therefore, the other two lighting categories were mostly focused on when analyzing clustering and results. An interaction plot (see Appendix Figure A2) was also displayed to check any initial standout effects on the data; there were no major outliers in these initial stages. The K-means clustering is done to group certain buildings that have similar trends in their lighting types. This can give a more in-depth look at each group the algorithm is composed of. To find how many clusters were needed for this analysis, the elbow method was used and the graph below is what was produced.

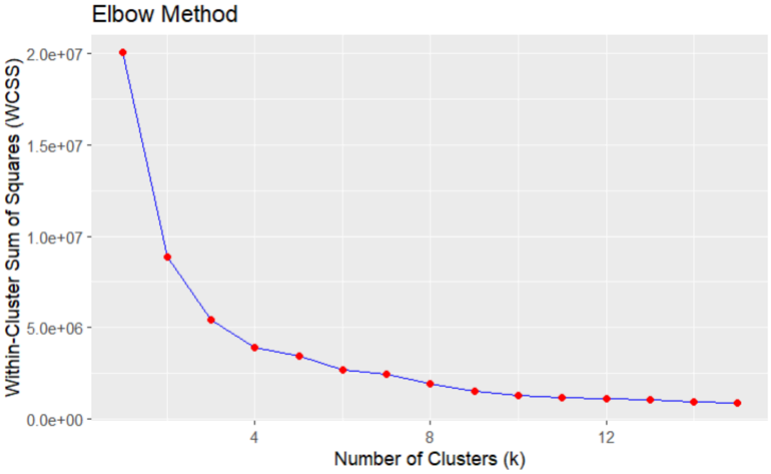


Figure 1: The graph above is a visualization of the elbow method to get the accurate number of clusters that should be used in k-means clustering.

From Figure 1, it can be seen that 4 clusters would be the best option. This is because after the step downward slope of the graph, at around k=4 the graph has a more constant slope. This is called the point where the graph forms the elbow-shape. Using this cluster number, we can group the data into 4 and get the output graphs below.

Results

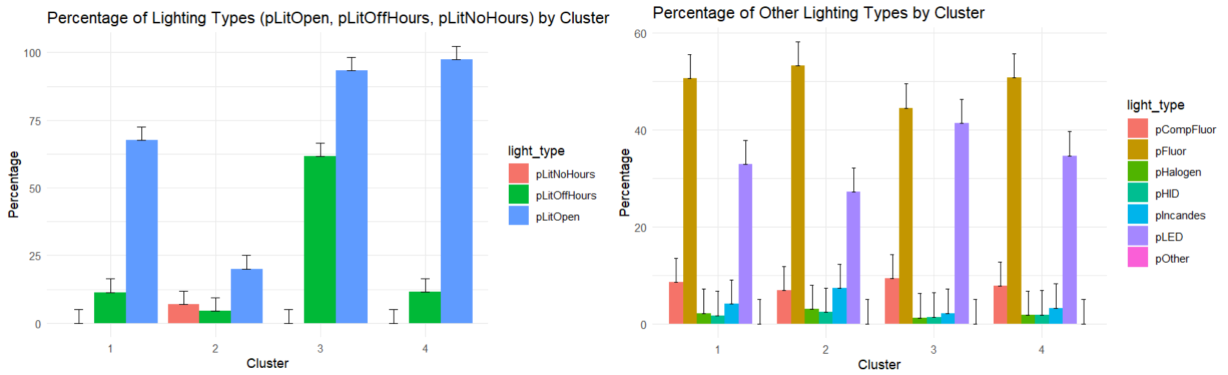


Figure 2: These two graphs visualize the 4 groups made by the k-means clustering. The left bar chart shows the differences in lighting percentages during times of the work day. On the right are the percentages of light types according to each cluster. Error bars are included in black.

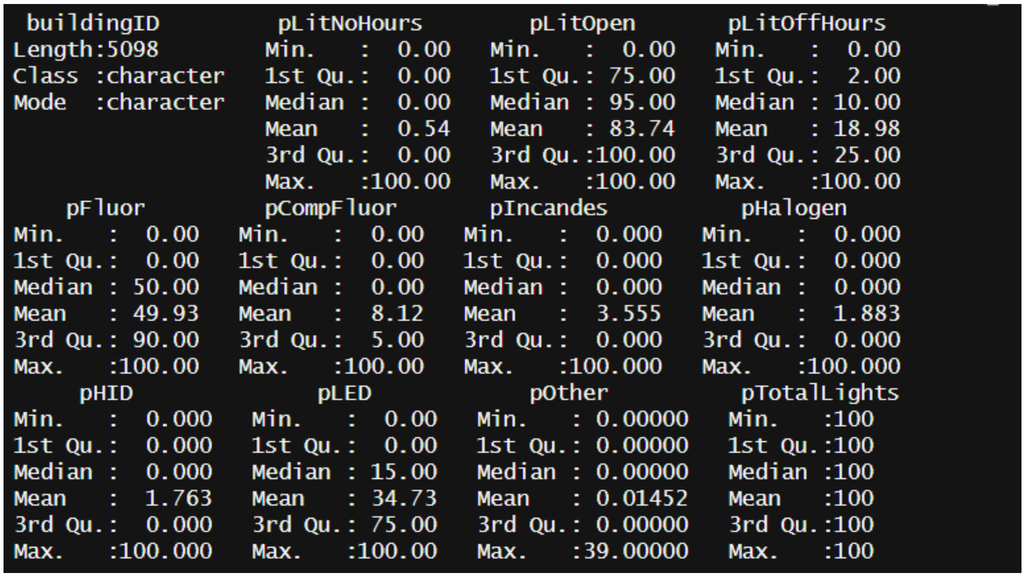
The number of data points for each cluster were 1173, 402, 820, and 2703 for clusters 1, 2, 3, and 4 respectively. We can observe from Figure 2 that cluster 1 and 4 are fairly similar, with cluster 1 having about 30% less percentage of lights lit when open compared to cluster 4. Cluster 2 was composed of most lights not kept on during any hours. Cluster 3 had the most percentage of lights kept on during the off hours. When looking at the lighting type data in the bar chart to the right in Figure 2, it can be seen that all the clusters are mostly fluorescent and LED lights. All the other lighting types seem to follow similar patterns across all the clusters. Exact mean calculations for each cluster can be found in the appendix (see Figure A3). The R code and dataset information can be found in the repository linked in the appendix.

Conclusion

Upon observing the bar charts in Graph 2, it is clear that these percentages lit during no hours, off hours, and open hours can be distinctly grouped in clusters. However, when looking at the lighting type percentages for each cluster there are no obvious differences between each cluster. Therefore, according to the data given by k-means clustering, there is not a difference in percentage of lighting types based on how often lights are kept on during various hours.

Appendix

Further graphs, tables, and links



```
 buildingID          pLitNoHours        pLitOpen          pLitOffHours
 Length:5098        Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
 Class :character   1st Qu.:  0.00   1st Qu.: 75.00   1st Qu.:  2.00
 Mode  :character   Median :  0.00   Median : 95.00   Median : 10.00
                    Mean   :  0.54   Mean   : 83.74   Mean   : 18.98
                    3rd Qu.:  0.00   3rd Qu.:100.00   3rd Qu.: 25.00
                    Max.   :100.00   Max.   :100.00   Max.   :100.00
    pFluor          pCompFluor        pIncandes          pHalogen
 Min.   :  0.00   Min.   :  0.00   Min.   :  0.000   Min.   :  0.000
 1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:  0.000
 Median : 50.00   Median :  0.00   Median :  0.000   Median :  0.000
 Mean   : 49.93   Mean   :  8.12   Mean   :  3.555   Mean   :  1.883
 3rd Qu.: 90.00   3rd Qu.:  5.00   3rd Qu.:  0.000   3rd Qu.:  0.000
 Max.   :100.00   Max.   :100.00   Max.   :100.000   Max.   :100.000
      pHID             pLED             pOther           pTotalLights
 Min.   :  0.000   Min.   :  0.00   Min.   : 0.00000   Min.   :100
 1st Qu.:  0.000   1st Qu.:  0.00   1st Qu.: 0.00000   1st Qu.:100
 Median :  0.000   Median : 15.00   Median : 0.00000   Median :100
 Mean   :  1.763   Mean   : 34.73   Mean   : 0.01452   Mean   :100
 3rd Qu.:  0.000   3rd Qu.: 75.00   3rd Qu.: 0.00000   3rd Qu.:100
 Max.   :100.000   Max.   :100.00   Max.   :39.00000   Max.   :100
```

Figure A1: Summary statistics for all variables in the dataset used for analysis.

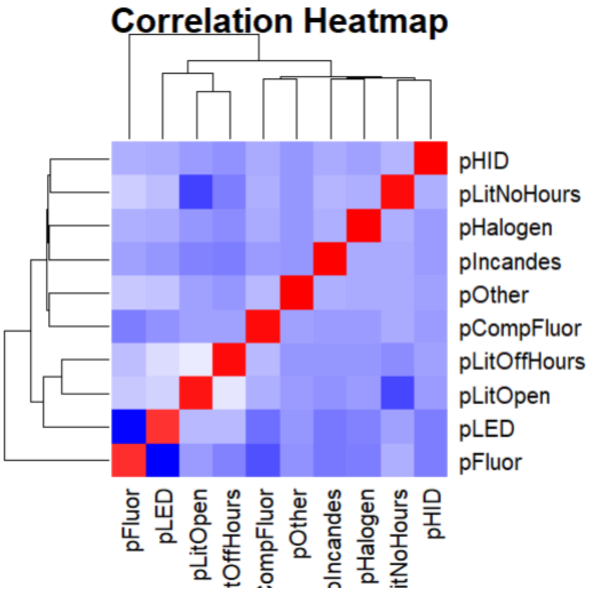
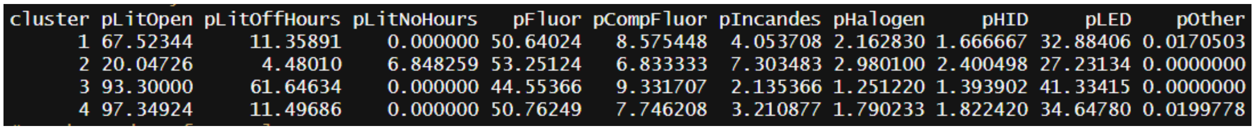


Figure A2: Correlation heatmap for all variables, more red indicates positive correlation, more blue indicates negative correlation.



```
cluster pLitOpen pLitOffHours pLitNoHours   pFluor pCompFluor pIncandes pHalogen     pHID     pLED    pOther
      1 67.52344     11.35891    0.000000 50.64024   8.575448  4.053708 2.162830 1.666667 32.88406 0.0170503
      2 20.04726      4.48010    6.848259 53.25124   6.833333  7.303483 2.980100 2.400498 27.23134 0.0000000
      3 93.30000     61.64634    0.000000 44.55366   9.331707  2.135366 1.251220 1.393902 41.33415 0.0000000
      4 97.34924     11.49686    0.000000 50.76249   7.746208  3.210877 1.790233 1.822420 34.64780 0.0199778
```

Figure A3: Table with mean calculations for all variables in all four clusters.

Link To GitHub Repository:
https://github.com/heleyna-tuck/Stats506_Project