**Problem:** Using Data Science techniques to determine in which neighborhood of the city of Halifax (Nova Scotia, Canada) to live. **Audience:** Families with young or adolescent children planning to live or relocate in Halifax.

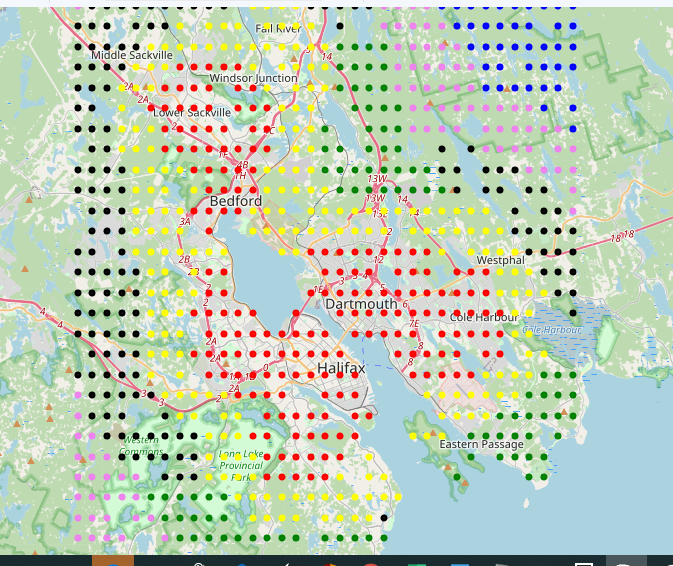
**Data:** The project consists in clustering the Halifax Regional Municipality (in Nova Scotia, Canada) areas by habitability type according to 6 parameters obtained from ‘’Foursquare’’: distance from an elementary school, distance from a Middle school, distance from a high school, distance from a park, distance from a hockey arena, and distance from a grocery store.

**Methodology:** A latitude and longitude grid was superimposed on the Halifax Regional Municipality (HRM) map. The grid points landing on a water plane were eliminated using the ‘’On Water’’ API. All the Elementary schools, Middle schools, High schools, hockey arenas, parks and grocery stores inside the grid area were extracted with coordinates using the Foursquare API. Then, for each point on the grid, the distance to the nearest venue by category was calculated. Six distances for each grid point were stored and those variables were used for clustering. For clustering, k=6 clusters were arbitrarily used (since there are 6 variables). The ‘’k-means’’ algorithm was used for clustering. The six mean distances were calculated for each cluster of coordinates.

**Results:** The first map displaying the clustered coordinates is presented in figure 1. Table 1 presents, for each cluster, the average distance and the rank (from closest to farthest) for each venue. Clearly, cluster 1 (red) is the one closest to all 6 venues wile cluster 2 (blue) is the farthest from all 6 venues.

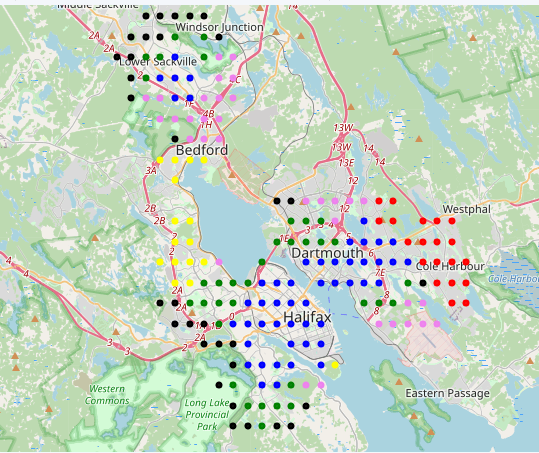
Since we do not obtain a lot of discrimination across venues, the coordinates in cluster 1 are solely considered in a second iteration of clustering. The map is presented in figure 2 and, the average distances and rank by new cluster, in table 2.

Cluster 9 (blue) is the one, overall, closest to all services, ranking first or second for all venues. It is not a surprise that cluster 9 (blue) corresponds to central areas of the Halifax peninsula, Dartmouth and Lower Sackville.

**Figure 1.** HRM map with first clustering (Black= cluster 0, red= cluster 1, blue= cluster 2, yellow= cluster 3, green= cluster 4, and violet= cluster 5).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Cluster** | **Elementary school** | **Middle school** | **High school** | **Park** | **Hockey arena** | **Grocery store** |
| 0 (black) | 6.28 / 5 | 5.03 / 3 | 4.45 / 3 | 6.71 / 3 | 8.53 / 4 | 4.61 / 3 |
| 1 (red) | 2.06 / 1 | 2.14 / 1 | 1.74 / 1 | 1.64 / 1 | 2.36 / 1 | 1.50 / 1 |
| 2 (blue) | 10.60 / 6 | 11.37 / 6 | 11.45 / 6 | 13.73 / 6 | 15.35 / 6 | 11.12 / 6 |
| 3 (yellow) | 3.46 / 3 | 4.14 / 2 | 2.99 / 2 | 4.05 / 2 | 4.82 / 2 | 3.23 / 2 |
| 4 (green) | 3.13 / 2 | 7.37 / 4 | 5.30 / 4 | 6.99 / 4 | 6.21 / 3 | 6.18 / 4 |
| 5 (violet) | 5.40 / 4 | 8.44/ 5 | 8.60 / 5 | 10.30 / 5 | 11.32 / 5 | 7.30 / 5 |

**Table 1.** Average distance (in Km) / rank (from closest to farthest from venue) for each cluster and venue type, first clustering.



**Figure 2.** HRM map with second clustering (Black= cluster 7, red= cluster 8, blue= cluster 9, yellow= cluster 10, green= cluster 11, and violet= cluster 12).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Cluster** | **Elementary school** | **Middle school** | **High school** | **Park** | **Hockey arena** | **Grocery store** |
| 7 (black) | 3.68 / 6 | 2.13 / 4 | 1.79 / 4 | 2.22 / 4 | 2.95 / 4 | 2.09 / 5 |
| 8 (red) | 1.20 / 2 | 1.74 / 2 | 1.27 / 1 | 2.69 / 6 | 4.63 / 6 | 1.34 / 3 |
| 9 (blue) | 1.06 / 1 | 1.25 / 1 | 1.33 / 2 | 0.96 / 2 | 1.64 / 2 | 0.90 / 1 |
| 10 (yellow) | 1.49 / 3 | 4.73 / 6 | 1.88 / 5 | 0.87 / 1 | 3.77 / 5 | 1.10 / 2 |
| 11 (green) | 2.67 / 5 | 1.90 / 3 | 1.35 / 3 | 1.24 / 3 | 1.30 / 1 | 1.36 / 4 |
| 12 (violet) | 2.13 / 4 | 2.73 / 5 | 2.97 / 6 | 2.27 / 5 | 1.94 / 3 | 2.29 / 6 |

**Table 2.** Average distance (in Km) / rank (from closest to farthest from venue) for each cluster and venue type, second clustering.

**Discussion:** Depending on the priorities of a family planning to relocate in the HRM region, this tool could be useful for selecting a neighborhood to settle in. For example, if all 6 variables are equally important, cluster 9 (blue) would be the preferred choice. If the distance to Hockey Arena is the most important variable, cluster 11 (green) would also be an option, and so on.

An alternative approach would have been, instead of using a grid, to use actual locations of houses on sale (if that data is available). Also, more thought could be put into selecting the number of clusters for the analysis according to clustering theory.

**Conclusion:** Clustering of the Halifax Municipal Region was performed according to a set of distances to nearest venues. The aim was to provide a tool to young families with children to select an area to relocate in the region. Clusters of coordinates nearest to important venues for families were identified.