The Battte of the Neighborhoods, Week 1

1. Problem description.

Background:

Hartford is the capital city of the state of Connecticut in the United States and is one of the most populated city in Connecticut. Hartford is an important center for medical services, research and education.

Although there has been some decline in the population in the last years, according to local and state authorities, several initiatives have been implemented by local authorities to create more opportunities for the fostering of economic growth. In this context, some commercial companies that are already operating in the nearby region, are assessing the possibility to expand its activities.

Business problem:

Having recognized an interesting opportunity to set presence in this geographic sector, a wholesale distributor, is assessing the convenience of buying property, and set operational hubs to distribute to retailers. In the company's experience, they consider the transportation cost as a major issue for its competitiveness, and of course, they have some restrictions in the amount of money to be invested in the acquisition of the property.

To find some guidance for this issue, we considered three metrics as the most important, to evaluate this possible expansion:

- Price of the property.
- Proximity to the principal train station, as this is the main means of transportation employed.
- Proximity to the possible retailers to whom we expect to distribute to, as this is the main driver for distribution costs.

So, the insights provided by this analysis should help to answer the question: how could we evaluate which geographic location is better for acquire a property to set our distribution warehouses up?

Target audience:

The solution to the posed problem is crucial to the executive level of the wholesale company, considering that these people will base their decision on where to acquire the property. Other possible stakeholer that can find it interesting, is the logistics management, for having better insights for costs estmation. Finally, it can be valuable for authorities to have a better understanding on availability of commercial venues by geographic area.

2. Data requirements.

Price of real estate:

Data of state of Connecticut real estate transactions was extracted from https://catalog.data.gov/, in csv format. As this data is detailed for the whole state and for all types of properties, additional processing is needed to limit the data for only transactions carried out in the Hartford county and for ignoring the transactions corresponding to residential properties.

Zip codes and geographical coordinates for each town/city in Hartford:

The business problem is about selecting geographic locations, consequently this data will be necessary but it is not in the real estate data set, but the zip-code can be used to get the coordinates. Neither there is available a single data frame with all cities in the county, so we extracted all zip codes and its related coordinates, by merging two tables:

- All cities with its zip codes, scraped from https://www.zipcodestogo.com.
- All cities with its geographic coordinates, in json format through an API call to https://public.opendatasoft.com/api/

By performing the necessary transformations, the data set with all transactions and geographic coordinates is obtained.

Venues of interest within the analyzed area.

It is important to remember that one of the important metrics explained explained, is the proximity to a selection of specific retailers (filtered by its category), to whom our items will be distributed. This can be solved by making a call to Foursquare API, to get all venues near each possible city, where we might establish the warehouses. With the call, we get up to 200 venues in a 5 km radius.

Final data set for analysis

After performing the needed processing and merging, the data set with all relevant data is produced. We will run a clustering algorithm, to get possible classifications based on the thee important metrics, shown in the three last columns.

	City	Zip Code	longitude	latitude	SaleAmount	StatDistance	Venue Counts	Labels
0	Avon	06001	-72.86431	41.789698	1800000.0	65.182714	18	2
1	Avon	06001	-72.86431	41.789698	600000.0	65.182714	18	2
2	Avon	06001	-72.86431	41.789698	2750000.0	65.182714	18	2
3	Avon	06001	-72.86431	41.789698	245000.0	65.182714	18	2
4	Avon	06001	-72.86431	41.789698	3300000.0	65.182714	18	2
5	Avon	06001	-72.86431	41.789698	600000.0	65.182714	18	2
6	Avon	06001	-72.86431	41.789698	825000.0	65.182714	18	2
7	Bloomfield	06002	-72.72642	41.832798	595000.0	65.044852	11	1
8	Bloomfield	06002	-72.72642	41.832798	500000.0	65.044852	11	1
9	Bloomfield	06002	-72.72642	41.832798	960000.0	65.044852	11	1