The Battle of the Neighborhoods, Week 2

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 stakeholder that can find it interesting, is the logistics management, for having better insights for costs estimation. 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:

The business problem is about evaluating options to acquire real estate property in Hartford, Connecticut. For this reason the most important metric for getting an answer is to have historical sales amounts, so data for transactions in the state of Connecticut was extracted from https://catalog.data.gov/, available in csv format.

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. Geographic coordinates data, related to zip codes was also necessary.

Both data sets were obtained from 4 different sources:

- 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/
- Prices of real estate property in the area of interest, in csv format downloaded from catalog.data.gov
- Venues to whom our items will be distributed, extracted in json format from foursquare.com, through an API call

3. Methodology

Exploratory analysis and data preparation

Having a good understanding of the business problem, it was concluded that we assumed the most important cost drivers for transportation. So, as we revised in detail the raw data extracted for the solution, the necessity of making transformations for each data set was revealed. For this reason, the exploratory and the preparation were carried out simultaneously.

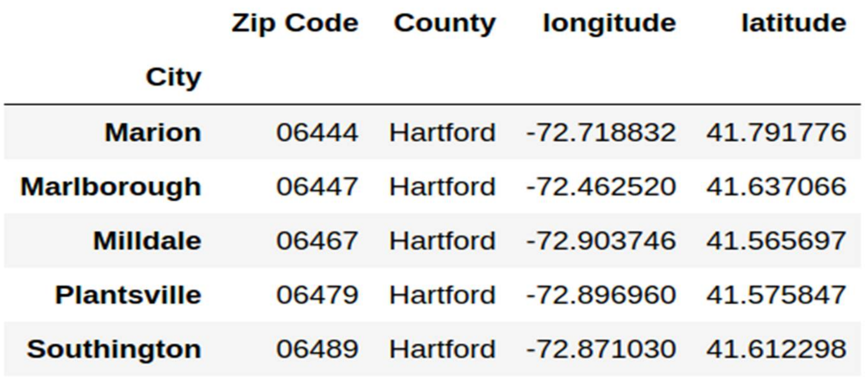
Zip codes and geographical coordinates preparation - ‘hartford_geo’ data frame:

Two tables were used in this step:

- All cities with its geographic coordinates, information gathered by making a call to an API. Iterative transformations had to be performed, because the relevant data was in nested dictionaries, as it is usual for json files.

- All cities with its zip codes, that was scrapped from a web page. Since it was in proper format, some the columns were renamed, all rows not corresponding to Hartford county were dropped, and the unique zip code were set as index.

Having prepared the tables, they were merged (using an inner merge) into a single one, using the zip code as the merging key. After this, only the records for Hartford county are included using a filter. Finally, the field containing the city (unique) has been set as index.



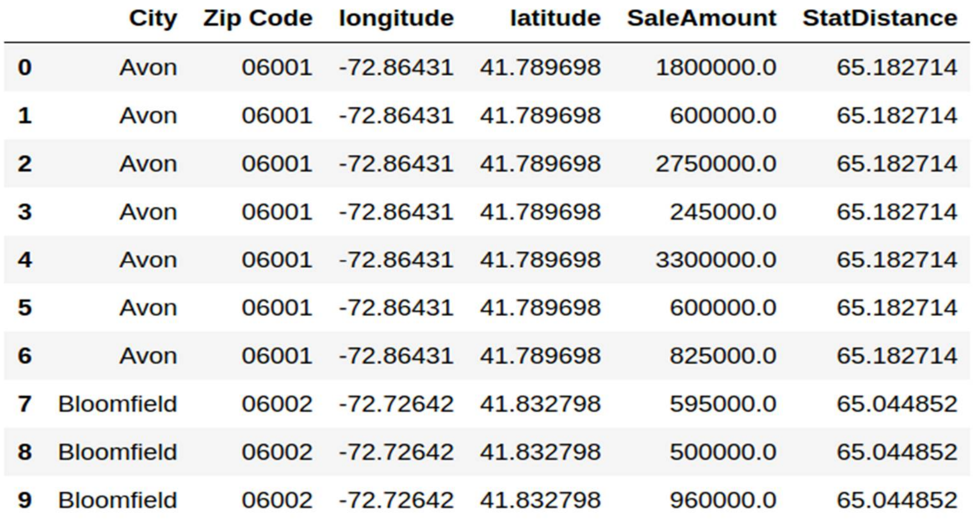
	Zip Code	County	longitude	latitude
City				
Marion	06444	Hartford	-72.718832	41.791776
Marlborough	06447	Hartford	-72.462520	41.637066
Milldale	06467	Hartford	-72.903746	41.565697
Plantsville	06479	Hartford	-72.896960	41.575847
Southington	06489	Hartford	-72.871030	41.612298

Price of real estate data set preparation - ‘cityprices’ data frame:

In the Data Requirements section, the relevance of prices of real estate was explained. However, the data extracted is detailed for the whole state and for all types of properties, being necessary to perform additional processing to limit the data to filter out the transactions corresponding to residential properties and to consider only transactions occurred later year 2015, thus having a good proxy for recent transaction amounts. For this purpose, boolean masks were used.

This table was then merged to the one containing the coordinates for each city, on ‘City’ column. In order to ensure that each city was included, even if it weren’t transactions recorded for that city, a left outer merge was run.

After the merge, an additional column –StatDistance– was included, including the proximity measure from any city to the main station. Here, it is important to remember that this metric importance was explained in the Business Problem section.



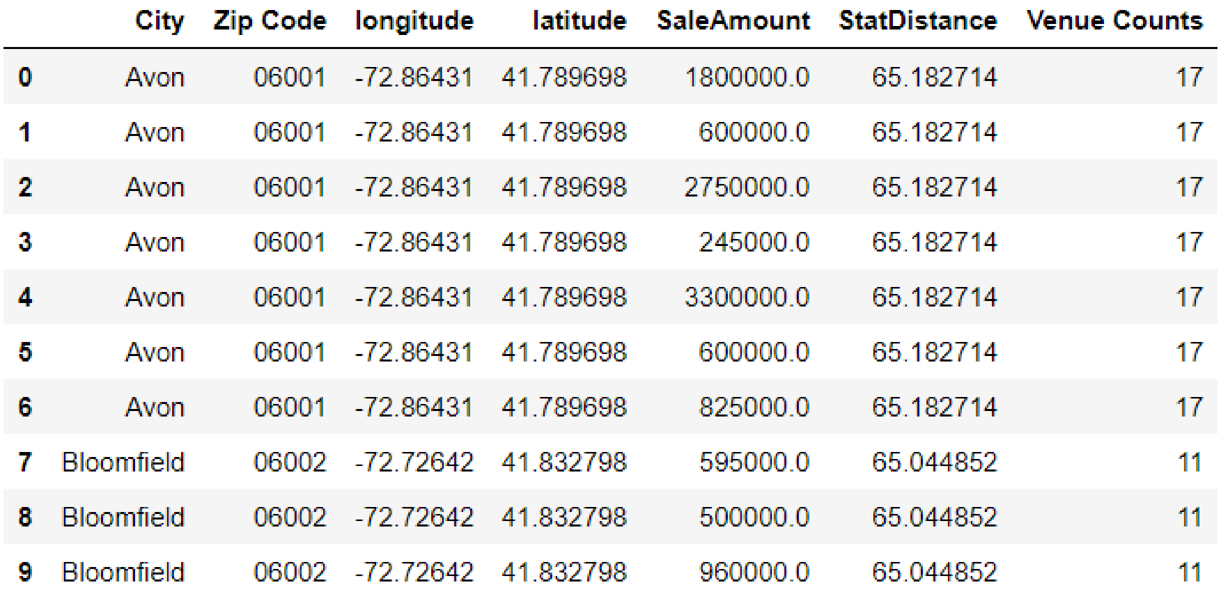
	City	Zip Code	longitude	latitude	SaleAmount	StatDistance
0	Avon	06001	-72.86431	41.789698	1800000.0	65.182714
1	Avon	06001	-72.86431	41.789698	600000.0	65.182714
2	Avon	06001	-72.86431	41.789698	2750000.0	65.182714
3	Avon	06001	-72.86431	41.789698	245000.0	65.182714
4	Avon	06001	-72.86431	41.789698	3300000.0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

Venues of interest within the analyzed area.

It is important to remember that one of the important metrics 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 – 'cityprices1' data frame:

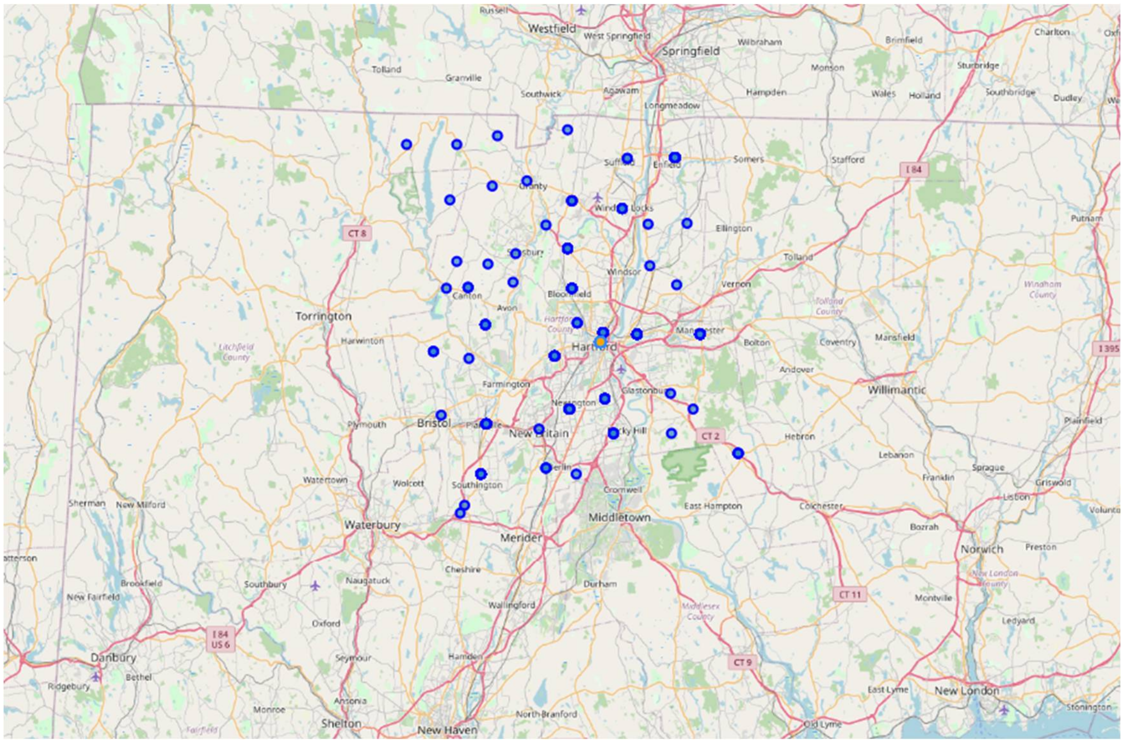
After performing the needed processing and merging, the data set with all relevant data is produced. We will run a clustering algorithm, to get possible classifications based on the thee important metrics, shown in the three last columns.



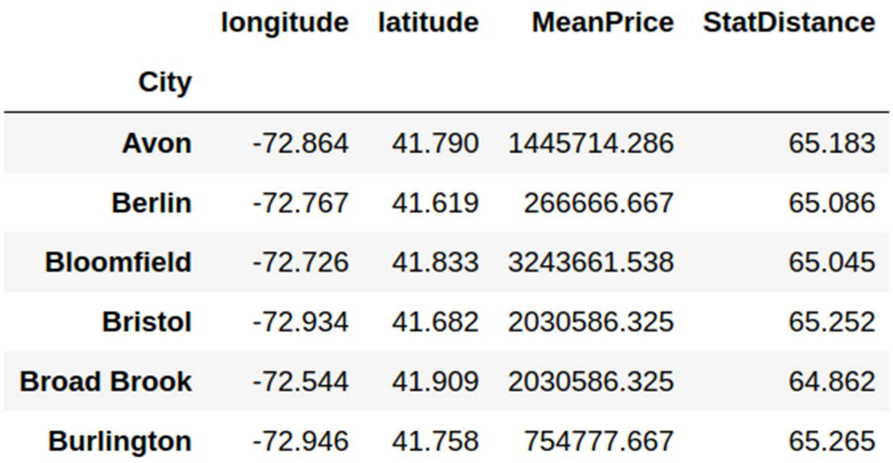
	City	Zip Code	longitude	latitude	SaleAmount	StatDistance	Venue Counts
0	Avon	06001	-72.86431	41.789698	1800000.0	65.182714	17
1	Avon	06001	-72.86431	41.789698	600000.0	65.182714	17
2	Avon	06001	-72.86431	41.789698	2750000.0	65.182714	17
3	Avon	06001	-72.86431	41.789698	245000.0	65.182714	17
4	Avon	06001	-72.86431	41.789698	3300000.0	65.182714	17
5	Avon	06001	-72.86431	41.789698	600000.0	65.182714	17
6	Avon	06001	-72.86431	41.789698	825000.0	65.182714	17
7	Bloomfield	06002	-72.72642	41.832798	595000.0	65.044852	11
8	Bloomfield	06002	-72.72642	41.832798	500000.0	65.044852	11
9	Bloomfield	06002	-72.72642	41.832798	960000.0	65.044852	11

Visual assessment of prices and geographic locations

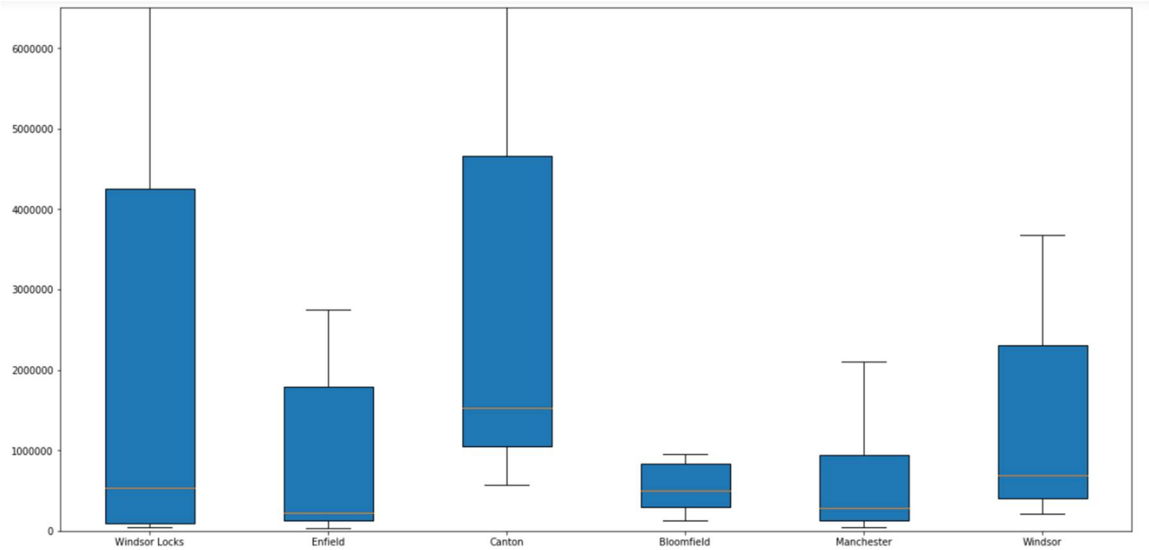
The point of all this analysis is about location. So, we considered important to include a visual tool to estimate how dispersed the possible locations are.



Whereas the price of real estate property is the most important factor for making any decision, and the assumption of meaningful differences between location and price, some statistic evaluations and visual support for making this point was necessary. So, we took the top 6 places, according to price and computed its means and generated box plots, to show how average prices differ between different towns within the area of interest.



	longitude	latitude	MeanPrice	StatDistance
City				
Avon	-72.864	41.790	1445714.286	65.183
Berlin	-72.767	41.619	266666.667	65.086
Bloomfield	-72.726	41.833	3243661.538	65.045
Bristol	-72.934	41.682	2030586.325	65.252
Broad Brook	-72.544	41.909	2030586.325	64.862
Burlington	-72.946	41.758	754777.667	65.265



It is possible to see the differences of prices, showing us that the ‘SaleAmount’ feature is very important and has influence in the business answer we are trying to answer. It also shows the presence of some outliers.

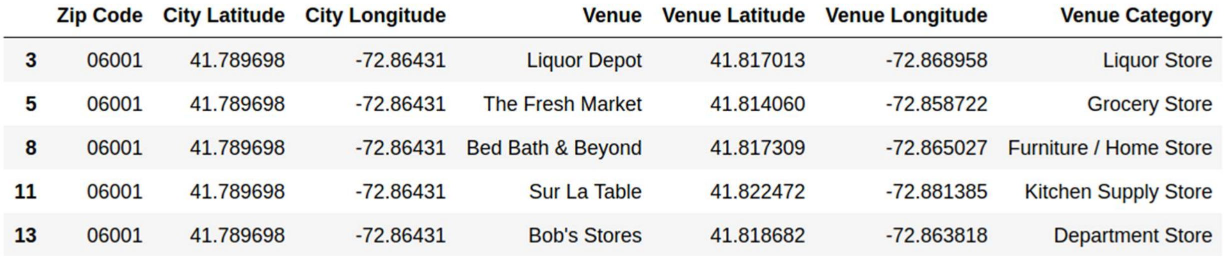
Retailers offering the products we distribute - ‘venueshartford’ data frame:

With all the meaningful cities and its coordinates, a request was made to Foursquare, to obtain up to 200 venues in a 5 km distance around each city. Since the results are in json format, the necessary transformations were done to get only the relevant information in a data frame.

Thereupon, we dropped all rows that corresponded to all venues that had no potential to have a commercial relationship with our company, irrelevant for the purpose of this analysis (isin method was employed). The list of venues included in the model is detailed below:

- Liquor Stores
- Grocery Stores
- Furniture / Home Stores
- Kitchen Supply Stores
- Department Stores
- Shopping Malls
- Deli / Bodegas
- Supermarkets
- Office Supplies Stores
- Convenience Stores
- Discount Stores
- Hardware Stores

- Miscellaneous Shops
- Warehouse Stores
- Markets



	Zip Code	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
3	06001	41.789698	-72.86431	Liquor Depot	41.817013	-72.868958	Liquor Store
5	06001	41.789698	-72.86431	The Fresh Market	41.814060	-72.858722	Grocery Store
8	06001	41.789698	-72.86431	Bed Bath & Beyond	41.817309	-72.865027	Furniture / Home Store
11	06001	41.789698	-72.86431	Sur La Table	41.822472	-72.881385	Kitchen Supply Store
13	06001	41.789698	-72.86431	Bob's Stores	41.818682	-72.863818	Department Store

Prices in Hartford county, with coordinates - 'cityprices1' data frame:

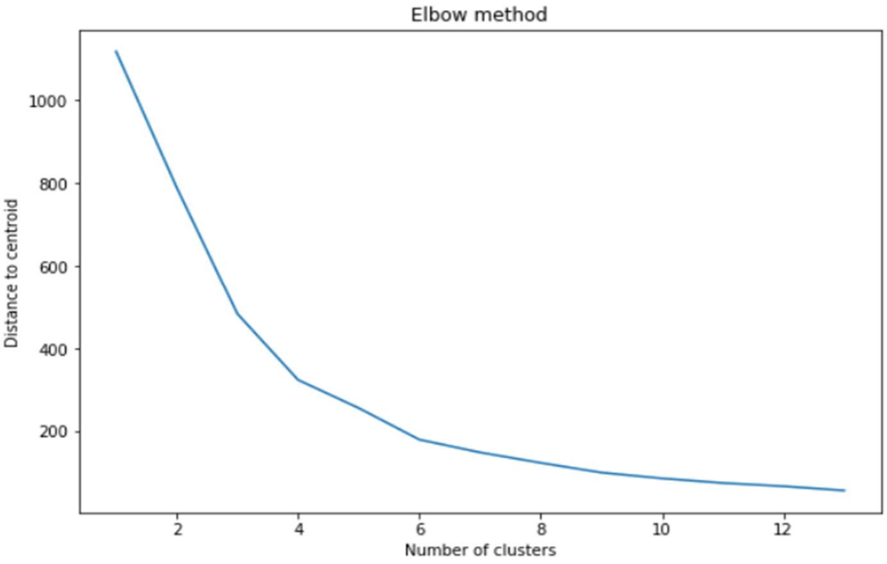
With all the external data procured, we merged all relevant data into a single table, containing all information related to geographic locations, real estate prices and presence of venues of interest. This was achieved by joining the last two tables, using the zip code as joining column.

4. Machine learning tools application – K-means clustering:

Having reaffirmed the importance of chosen metrics, we need a tool for reducing all transactions and locations data to some common categories, allowing us to decide which category is the best from the perspective of the distribution costs, the k-means clustering appears to be very adequate.

Choosing the best number of clusters and running the algorithm:

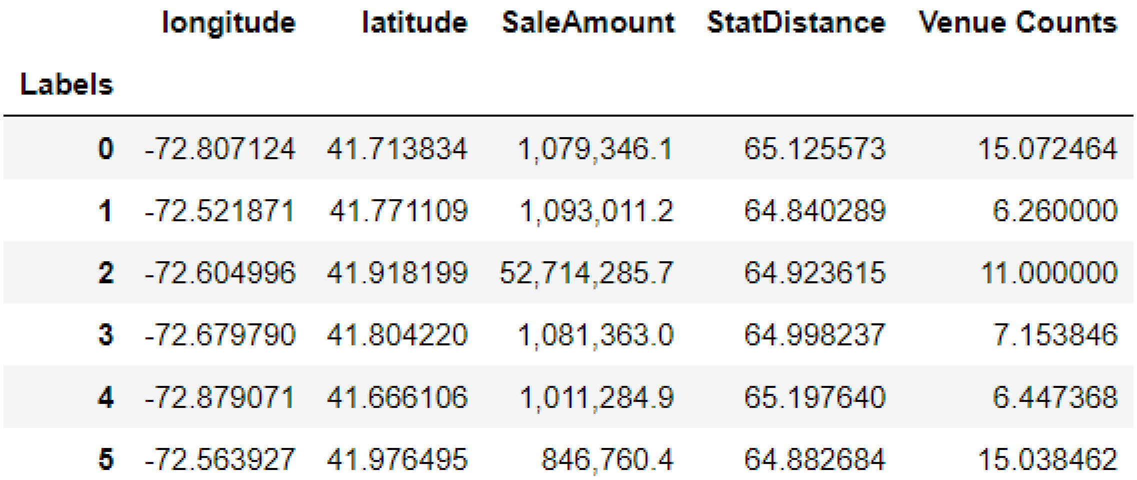
To know the adequate number of clusters for this problem, K-means clustering was run number of clusters ranging between 1 to 12 and a plot was produced showing cluster number versus the metric for the distance of observations to its centroids.



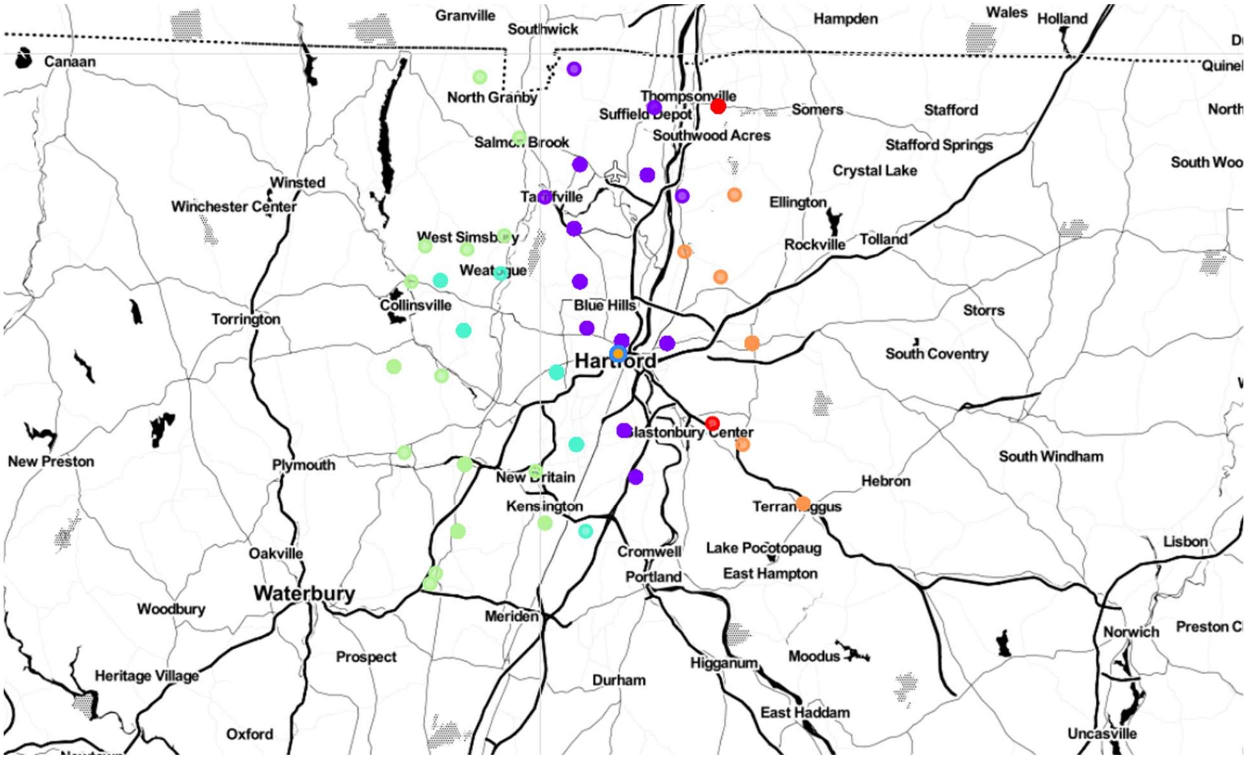
The graph shows that the distance to centroids measure starts decrease rate is lower starting in k=6, we considered this number of clusters to continue the analysis.

Results:

The resulting labels were included to the base data frame as a new column and the **mean values** in all relevant metrics were computed.



	longitude	latitude	SaleAmount	StatDistance	Venue Counts
Labels					
0	-72.807124	41.713834	1,079,346.1	65.125573	15.072464
1	-72.521871	41.771109	1,093,011.2	64.840289	6.260000
2	-72.604996	41.918199	52,714,285.7	64.923615	11.000000
3	-72.679790	41.804220	1,081,363.0	64.998237	7.153846
4	-72.879071	41.666106	1,011,284.9	65.197640	6.447368
5	-72.563927	41.976495	846,760.4	64.882684	15.038462



5. Discussion

Before giving meaning to the results shown, it's convenient to explain the criteria to evaluate each cluster: the best locations are the ones that have lower prices, their distances to main station is as low as possible and that have the maximum possible of venues of interest within the radius set (5 km). Given this criteria, a general evaluation of each cluster could be performed:

- **Clusters 1, 3 and 4,** which have reasonable prices their mean distance is around 65 units (still in coordinates) and with relatively low venues of interest in their proximity. It would not be advisable to acquire properties belonging to this groups, because there are other with the same levels of prices with more venues in their nearby and similar in proximity to train station.
- **Cluster 2,** that have properties with levels of prices that makes it impossible to even consider buying
- **Cluster 0,** that appears to have prices relatively reasonable, and somewhat far from the station, but still a good candidate since it has the higher number of venues near.
- **Cluster 5,** that has the best combinations of features according to our criteria.

As it is evident that Cluster that has better, respect to the election criteria previously defined, the correspondent regions should be preferred for an eventual acquisition. Even in the case of

Even when the results allow us to distinguish which the best options are for making the acquisition. However, some additional investigation could be done to extend our understanding on variation of prices for Cluster 2, which appears to have unusual high values. This could also be a good support for decision in the case of considering renting, as this analysis can give a good understanding of the locations with better combinations of the important metrics.

Although results appear to be very clear, it still has to be revised by the executive instances within the company and other related areas such as the ones in charge of commercial and logistics issues.

6. Conclusions

For this project, an analysis for giving a wholesale distribution company some guidance for deciding where to buy real estate property to establish operational hubs to deliver products to retailers. Geographic, transactional and geographic presence of specific businesses information was used, to be explored transformed and feed into a non-supervised classification model.

Results were somewhat clear, defining one specific cluster as more convenient in the case of making the acquisition, although all insights have to be analyzed by other knowledge areas in the company. This also could lead to make some adjustments or enrich the models.

It would be worthy to investigate in more detail some data sets, as outliers and values and dramatic differences in mean values have been identified.