IDENTIFYING INVESTMENT OPPORTUNITIES IN KANSAS CITY

INTRODUCTION

Investment is fundamentally a matter of identifying assets which can be purchased for less than their actual long-term worth. In this project we considered the problem of trying to locate an industrial concern in Kansas City, Missouri. Kansas City has a long history as an industrial hub, containing factories for Ford, Honeywell, and many other major manufacturing firms. Broadly the business environment of Kansas City is extremely favorable for industry, but to actually invest, stakeholders need more information to narrow down their search. We will be interested in identifying locations in the city which are business-friendly, but not yet heavily industrialized, soas to identify them as valuable areas for such investments.

We will be using cluster analysis to determine similarity between the business profiles of given areas. We would first like to distinguish between broadly residential areas and areas which house many consumer-oriented businesses and those which have a more specialized business environment. We can then further distinguish those which are most friendly to industry and those which are not already heavily developed for manufacturing. This will give us a good starting point for further investigation of an ideal location to place a new manufacturing business.

DATA

We analyzed an area within twelve kilometers of the center of the city, which assures a new business of access to the local labor market. We generated longitude/latitude pairs forming a hexagonal grid covering this area and query it with the Foursquare API to get lists of businesses local to each area. We then queried these points with the Foursquare API for local businesses and performed cluster analysis to sort each one of these small areas into a different cluster based on the local business environment.

Cleaning and formatting the data for analysis involved making one-hot vectors for each ‘venue’ returned from Foursquare and then grouping them by which neighborhood they corresponded to and getting their frequency. This gave us a dataframe that had

each neighborhood's business environment summed up in its columns, appropriate for doing K-means analysis.

We had to drop rows for which no businesses were returned. These correspoded, when mapped, almost entirely to rivers or other major geographical barriers, so they should not form any problem for our analysis. This brought our total number of 'neighborhoods' down from over 1400 to just over 1200, a noticeable reduction but not one which is in any way prohibitive.

METHODOLOGY

To determine K for K-means the 'elbow test' was attempted by graphing error against K, but the plot was linear, indicating that the fit gets increasingly better as more clusters are added. Having too many clusters is useless for decision-making, because then you have not succeeded at meaningfully reducing the quantity of data that needs to be considered for each location. 1200 neighborhoods is too many for a stakeholder to evaluate meaningfully, and so would a higher K delivering a lower error.

Instead of using the 'elbow test', a different approach was called for. By graphing a historgam of the number of neighborhoods in each cluster, we identified for which K new clusters tended to become very small, consisting of just one or a few neighborhoods. This happened at approximately K=6. Since too many clusters are only a problem if they become overwhelming, and too few could deny us insights we might have benefitted from, K-8 was selected as the best choice. Possibly this delivers more clusters than we really need, but smaller clusters can be meaningfully compared against their geographical neighbors if they do not have enough members to have a meaningful character, and so nothing is really lost by their inclusion provided there not too many of them for us to deal with.

Having obtained these cluster labels, we plotted them for each of our neighborhoods on a folium map.

RESULTS

Cluster analysis efficiently sorted our data into three very broad cateogories, from which we can then make finer distinctions. There are areas that return no results for nearby businesses, which are not clustered or mapped at all. These areas are, for either geographical or regulatory reasons, off limits. There are also areas corresponding to our

largest category, that where no particular category of business stands out, or our ‘close to zero frequency’ cluster. These correspond broadly to residential districts, and can also be ruled out as likely targets for industrial investment.

Our third broad category is all of our other clusters, which show a business environment distinct from the baseline. We can see these neighborhoods appearing in distinct groups across the city, and these form our areas for further consideration. By narrowing down even more precisely on the characteristics of these neighborhoods we can figure out which are the most, and which the least promising for our purposes.

DISCUSSION

Although all of our clusters show factories as a fairly common category, several stick out in particular. Cluster seven is definitely the most industrialized, encompassing the Fairfax Industrial District. This assures us that our analysis has the ability to distinguish the traits we are interested in; however, as the goal of this analysis is to identify cost-effective investments that might otherwise go unnoticed, we can exclude the industrial district itself from our consideration and look instead at areas that resemble it according to our clusters.

Other areas assigned to the same cluster as the Fairfax Industrial district appear in small blocks. One of these in the east, one in the northeast, and one in the southeast of the city. All three of these small areas of cluster seven are bordered by areas of other clusters which are similarly distinguished from the baseline cluster we want to exclude from consideration. Because they are in neither cluster and border areas in our industrial cluster, we can expect them to have a favorable environment.

There is a small area in our fourth cluster on the northeast that meets these criteria, and then another two areas in cluster one on the west and southwest. Because they form distinct blocks of business concerns that differ from baseline and they border industrial-friendly neighborhoods, these constitute promising avenues to explore for building the sort of business we are considering.

Last we can exclude our ‘cluster zero’. It shows a high prevalence of factories, indicating a heavy industrial prevalence. Much like the Fairfax Industrial District, this makes it an area that is already subject to heavy industry investment and therefore a less

cost-effective opportunity. Buying things cheaply is a cornerstone of good business practice, and areas that are already highly developed tend to be expensive.

CONCLUSION

Our goal was to provide a good starting point for stakeholders to identify opportunities for investment in the Kansas City area for industrial concerns. Ultimately stakeholders will have to make a decision based on the particulars of the specific business environment, including the price of real estate, local regulations, and ease of access for workers. However, this analysis has provided a good starting point for them to do further investigation.

This method also provides a good method of analyzing the business environment of different districts in any city; as the code is modular it can be generalized to any major industrial area, or used to compare them against each other. Areas for investment have a sort of ‘goldilocks’ criteria. They are both attractive enough to seem like viable opportunities and undervalued enough to be cost-effective. Using cluster analysis to quickly identify which areas meet both of these criteria is a useful solution to the problem of identifying this type of potential.