Comparing Neighborhoods for Restaurant Viability in Kansas City, Missouri

Landis Duff

1. Introduction

1.1 Background

Kansas City is the largest city in the US state of Missouri. It has many neighborhoods with varying properties, there are residential, industrial and retail areas. In this project we examine where is the best neighborhood to open a restaurant serving contemporary American cuisine.

1.2 Problem

It will be advantageous to analyze data in order to determine what the 'best' neighborhood to open this new restaurant. We will use data to determine where the city is most responsive to service requests in order to gauge which neighborhoods are well-kept by the city. Furthermore, we will check which neighborhoods have the highest crime rate focusing on crimes against persons and property crimes in order to consider the safety of each neighborhood. We will also look at which neighborhoods have similar restaurants to determine which neighborhoods would be most receptive to a new contemporary American restaurant.

2. Data Acquisition and Cleaning

2.1 Data sources

Data will be collected from Kansas City's open data website. This website has a database of 311 call center service requests to the city. We will also use data from Foursquare API to determine which restaurants already exist in each neighborhood. I also used, from Kansas City's open data site, crime data. I then used shapely to determine in which neighborhood each crime was committed in 2018. On KC's open data website they also provide a neighborhood boundary geoson which I relied on heavily.

2.2 Data Cleaning

Now, the data I collected was not formatted correctly for the analysis. So first I grouped the data in the 311 call center data by neighborhood and calculated the share of calls closed within the expected timeframe and included this, and the number of days to close as well as the total number of calls in the final dataframe.

The police crime data provided latitude and longitude information for each crime, but it did not tell us in which neighborhood each crime occurred. So in order to group the crime data by neighborhood we need to determine this. Now, the neighborhood geojson file contains a feature that defines the vertices of a polygon. I input this data as a polygon in shapely, then I used the shapely contain method to determine in which neighborhood each crime was committed.

Since the neighborhood data was given as polygons, I needed to calculate the centroid of each polygon in order to have one point associated with each neighborhood to check for nearby venues using Foursquare API, so I wrote the code to calculate the centroid of each neighborhood, then added this to the data frame.

Note that all of the data that I used came as individual data points with a column denoting the neighborhood, so I grouped all of the data by Neighborhood in order to create the data frame where I did my analysis.

Finally, I used Foursquare API to find similar venues within one kilometer of the centroid of each neighborhood. While this won't give the total number of each type of restaurant in each neighborhood, it will give us a solid picture of what is close.

3. Methodology

3.1 Exploratory Data Analysis

While exploring the data I looked at two features gleaned from the 311 data that looked valuable, namely, the share of calls that were completed in the estimated time frame and the average number of days it took to close each call. Now, I checked for a correlation between these two variables and there was a weak negative correlation, so generally the more days it took to close a call, the more likely it was to not be closed on time. However, this correlation was weak enough to consider these as separate features.

I also plotted each of the features on a choropleth map to get an idea of which neighborhoods had which features. Now, from these choropleths it was clear that neighborhoods near downtown had the highest number of American restaurants, and also had higher crime rates, however, what was less clear is what was

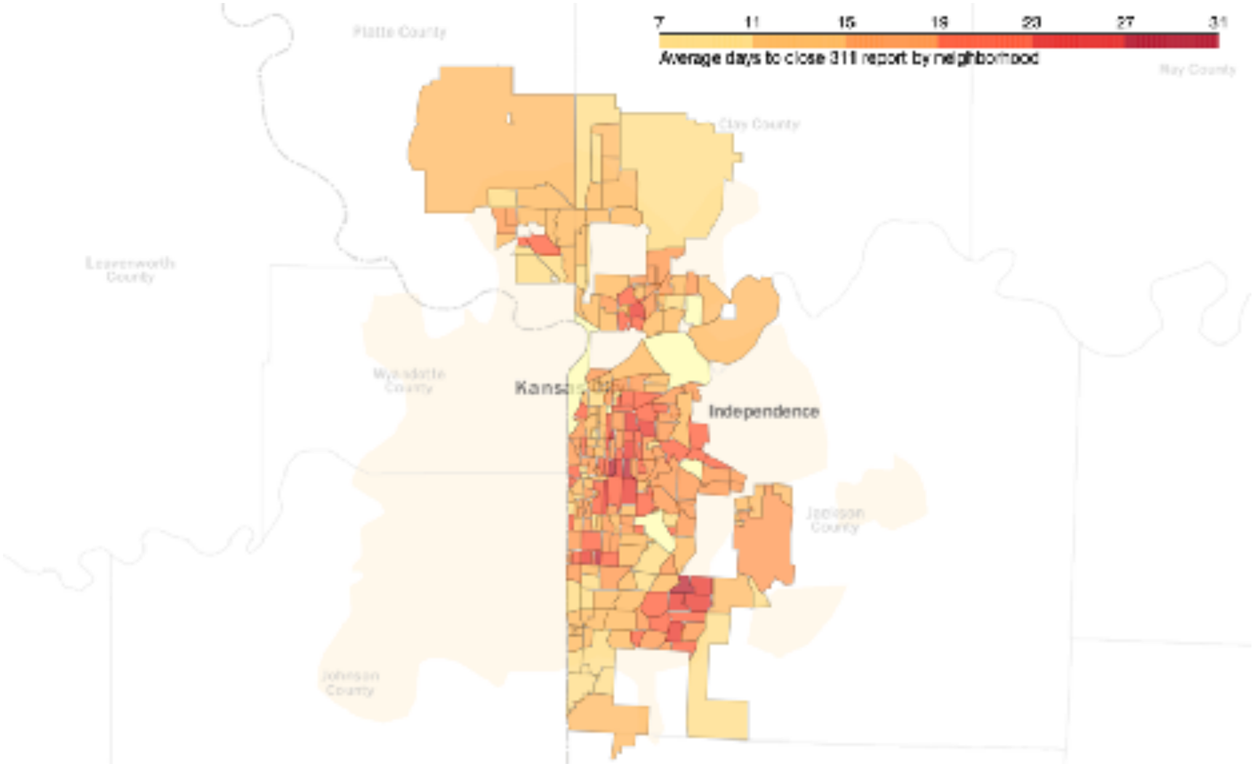

Figure 1. Average number of days to close a 311 call by neighborhood in Kansas City

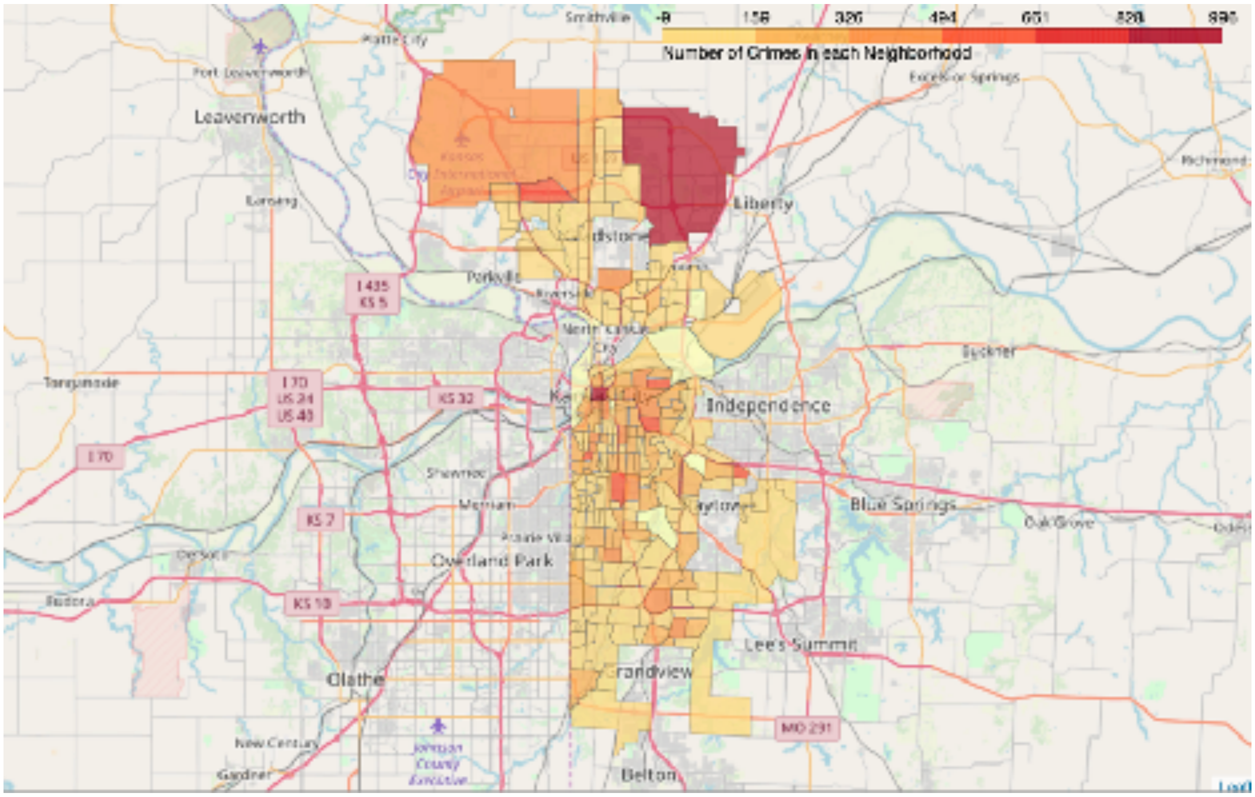

Figure 2. Total number of selected crimes in each neighborhood

going on with the 311 data. The amount of time it took to close calls appeared to be more randomly distributed.

3.2 Machine Learning

I used two different unsupervised machine learning algorithms to cluster the neighborhoods in Kansas City. First, I used k-means. I used the elbow method to determine what the ideal number of clusters was. Now, there was not a clear elbow, but there was definitely an inflection point at k=5, so we used 5 clusters for this algorithm. Recalling that k-means is a heuristic algorithm that relies on randomized centroids as the starting point I ran this algorithm a few times with a few different clustering results. Now, this tells me that the cluster structure was not particularly strong. Now, since we had four dimensional data, the clusters are not directly visualizable, but we did plot them on a map, using a marker on the neighborhood centroid to indicate by color which cluster each neighborhood was in.

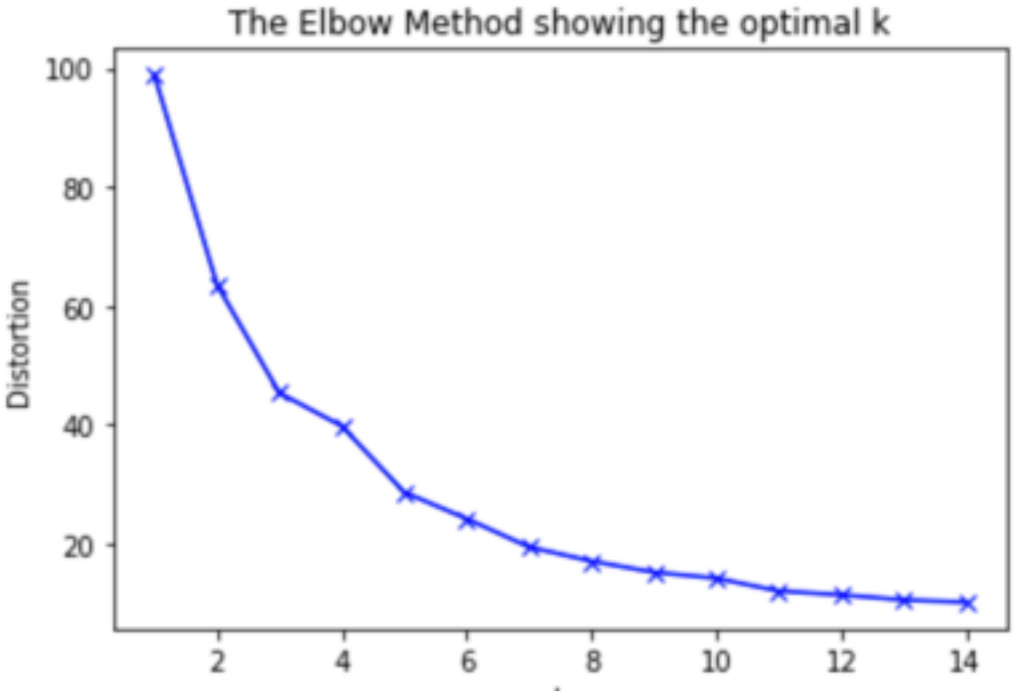

Figure 3. The elbow method for finding the optimal k.

Secondly, I used agglomerative clustering to cluster the data. Now, agglomerative clustering can be used without specifying a number of clusters, but in order to get the best distribution of clusters I chose seven as the number of clusters for the algorithm to use. Agglomerative clustering is not dependent on a randomized starting point, so if the same features and the same number of clusters are used each time we will always get the same clusters.

4. Results

I noticed in both sets of clustering that I did that the ideal clusters specified by the machine are neighborhoods surrounding the downtown area, and Country Club Plaza area which are both the main restaurant districts in Kansas City.

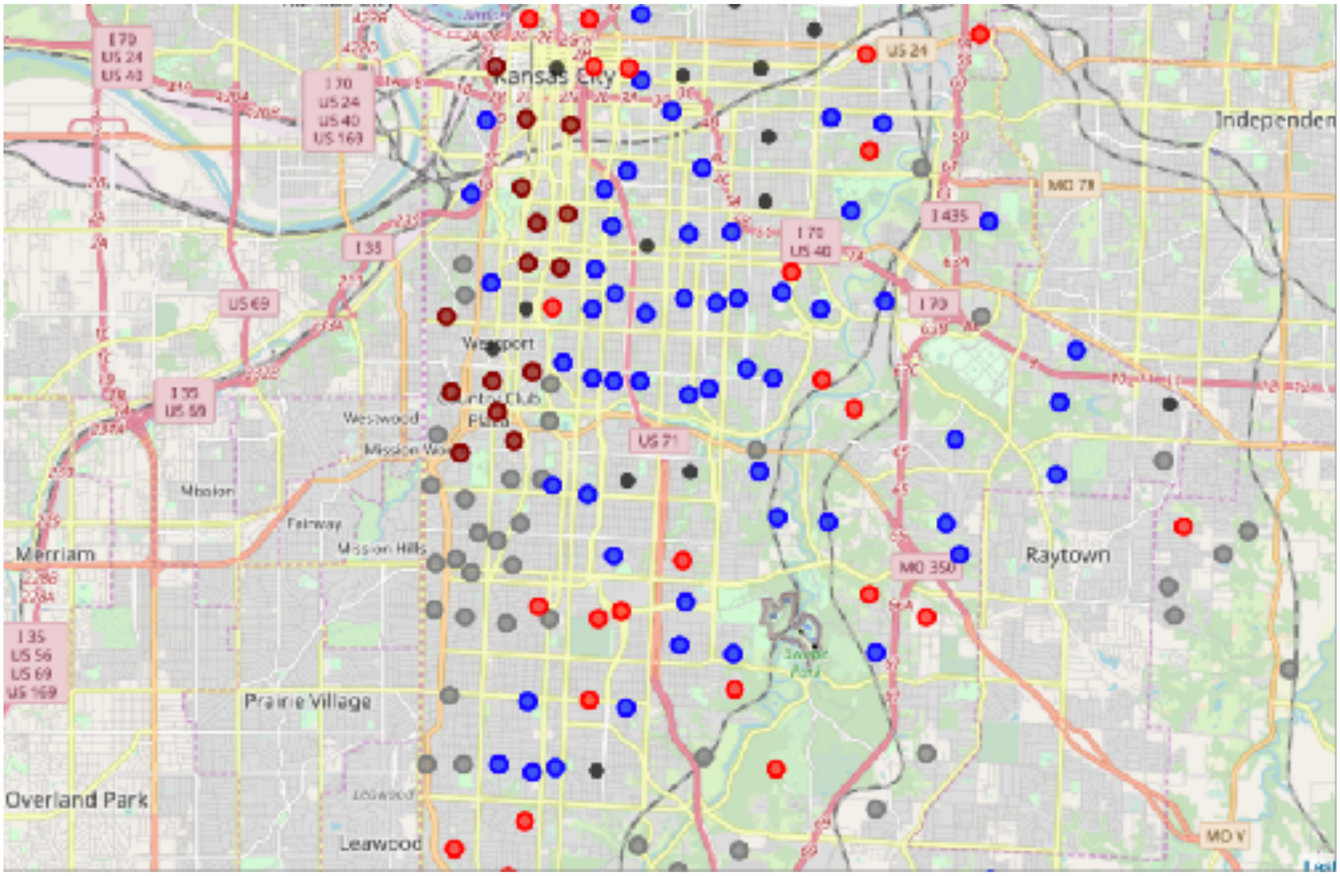

Figure 4. Neighborhoods south of the Missouri River clustered by k-means. The grey points specify the most ideal neighborhoods. The dark red are the least ideal since they have a high number of American Restaurants.

In figure 4 the grey neighborhoods are the most ideal neighborhoods from the k-means algorithm to open a new restaurant. Now, the blue neighborhoods are also favorable according to our metric, but these are primarily residential neighborhoods.

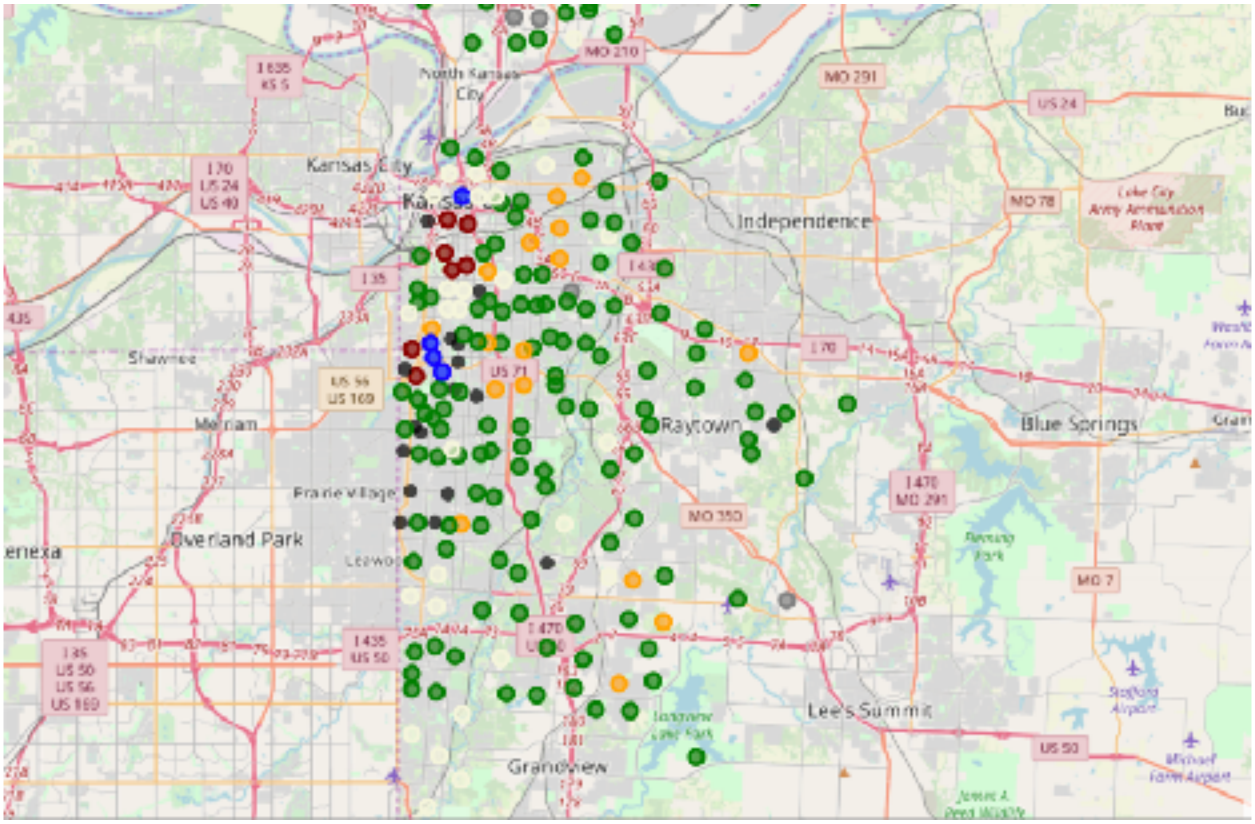

Figure 5. Agglomerative clustering of neighborhoods in Kansas City. The green and grey neighborhoods had the most desirable features in this clustering

Using agglomerative clustering, we did not get quite as much separation of neighborhoods as we did with k-means clustering. The green neighborhoods seem fairly ideal for this, although many are primarily residential.

5. Discussion

The analysis above shows that based on the factors we considered there are quite a few neighborhoods in Kansas City that would be good candidates for a new American Restaurant.
Based on my personal experience in Kansas City our agglomerative clustering produced neighborhoods that make more sense in our analysis. From the way that the program clustered it seems clear that the number of crimes and the number of American Restaurants had more weight in terms of the clusters than did the 311 data.

Perhaps we would have benefited from using min-max scaling in our data set. However, this depends on what factors we consider to be most important. Clearly if all factors have equal weight this would be a better approach, but I would argue that the number of American Restaurants currently near a neighborhood as well as the amount of crime in the neighborhood are likely more important to the viability of a business than the 311 data. However, the results that we got in agglomerative clustering show that the best place to build a New American Restaurant is in the neighborhoods around the major restaurant districts in Kansas City, (namely the Country Club Plaza and CBD Downtown Area). Now, from experience these neighborhoods are mainly residential, but those living there could perhaps benefit from having a restaurant closer to home.

A future improvement to this could include determining which neighborhoods are primarily residential versus commercial, and separating these neighborhoods. Generally commercial neighborhoods would be more ideal for opening a restaurant rather than primarily residential neighborhoods.

6. Conclusion

The purpose of this project was to identify which neighborhoods in Kansas City would be the most ideal to open a new American Restaurant. We used four factors to decide which would be the best. In examining the factors, it is possible we could have divided the number of crimes and the number of American Restaurants by the area of each neighborhood in order to more properly weight those factors. However, it was difficult to find the area of each neighborhood online. However, we did identify a range of neighborhoods that would be ideal for opening a new restaurant.

The final decision on where to open a new restaurant could include this analysis, as well as advice from those with more local business acumen as well as understanding of the neighborhood dynamics. Certainly, the stakeholders should take all of these factors into account when making their final decision.