

**Kansas City Postal Code Clustering**

Created by: Brandon Lombardino

Created Date: 05/12/2020

**Introduction:**

A business wants to know which areas of Kansas City are ideal to open a coffee shop. The owner hopes to draw in customers by building a unique relationship within the community rather than through low prices or innovative products. The report will investigate which area in Kansas City is best suited for the business given the owners goals and expectations

**Data and Sources:**

The data is based on postal codes from the Kansas City area. The postal codes are used to define regions within the Kansas City metropolitan area and have coordinates to map their locations. The data will be used with folium maps and a Scikit-Learn cluster model to segment the region based on venue information imported from Foursquare.

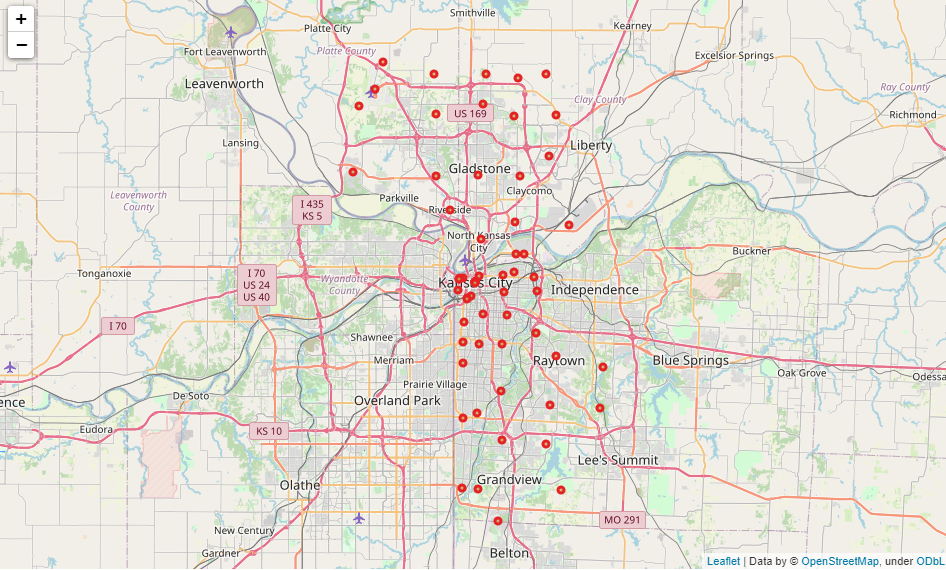
**Methodology:**

Collecting the Data

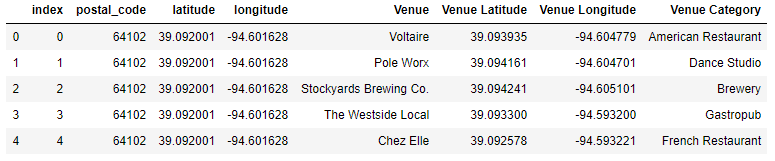
To start the analysis, a data table containing postal codes and coordinates needed to be imported. The Kansas City metropolitan area has 70 postal codes to include. The corresponding coordinates were obtained manually by copying the latitude and longitude data off public maps. After the data was collected on an excel file, the data table was transformed into a csv before being imported into Jupyter Notebooks. Python 3 was the language used within Jupyter Notebooks. After the data was imported, the geocoder package was used to obtain coordinates for Kansas City. Later in the analysis, Foursquare is used to import venue data related to the given postal codes. A free developer account on Foursquare.com is needed to obtain the Client ID, Client Secret, and version data needed to run an API to retrieve data from the external database. The venue data is limited to 500 venues within a 1,000-meter radius of the coordinates.

Preparing the Data

Using the coordinates, each postal region is mapped using the folium package. The map visualizes how the Kansas City area is dispersed.



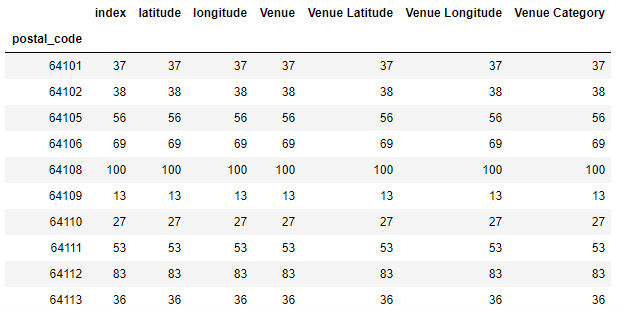
After the map is created the venue data from Foursquare is organized into its own table based on the venue name, category, latitude, and longitude. The postal code and venue data tables are then merged to create a dataset displaying postal codes, latitude, longitude, venue, venue latitude, venue longitude, and venue category.



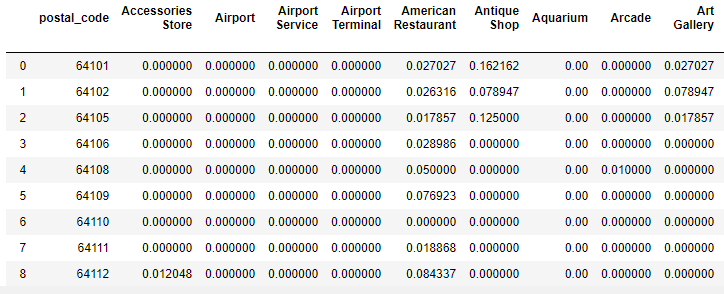
Later, more data preparation is needed to prepare for clustering. To optimize the clusters, the postal codes with less than 10 venues associated with them were dropped from the model. The postal codes with less than 10 venues were deemed insufficient to use in the predictive model. After 27 postal regions were dropped, the data was ready for one hot encoding, a feature engineering tool that transforms a categorical column into multiple dummy variable columns. The dummy variables are needed to quantify the data set as is required for the clustering model.

Exploratory Analysis

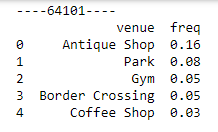
The first analysis created was a matrix that counted the number of venues by each postal code. It was after the venue count table that it became apparent that many postal regions needed to be eliminated from the model due to lack of venue data.



After grouping postal codes by venue counts, postal codes were again grouped, but this time by taking the mean of the frequency of occurrence of each category.



The mean of the frequency of occurrence of each category can then be used to display the top 5 most common venue categories in each postal region.

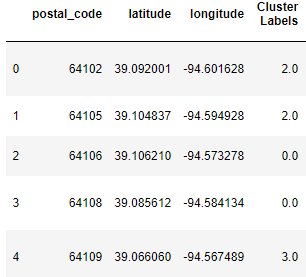


Clustering

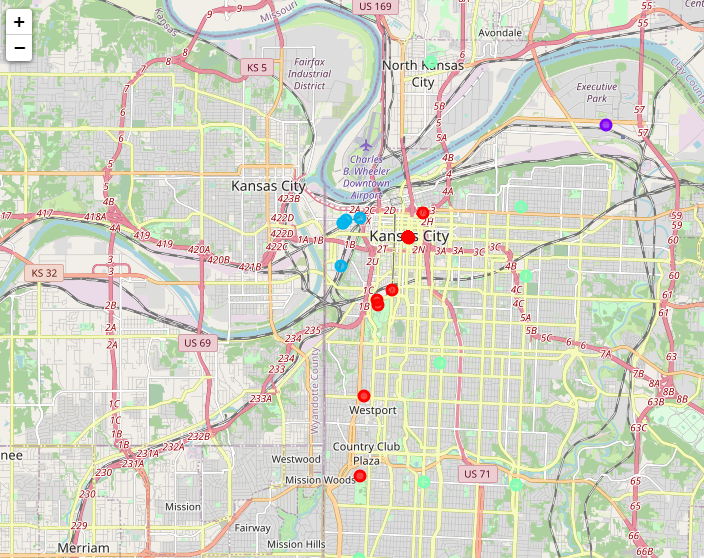
The clustering analysis begins by fining the top 10 most common venues for each postal code and displaying them within a dataframe.



The dataframe can then be used with a k-Means clustering model to show which postal regions are similar and different from one another. K-Means can segment the Kansas City area by dividing the data into non-overlapping clusters without any internal structure. After a series of trial and error, 5 cluster was deemed ideal for this analysis. Setting k = 5 means the model will use the given postal codes to split them up into 5 distinct clusters. The clustering analysis adds an additional column to the dataframe called “Cluster Labels” which contains integers between 0 and 4. In other words, the clustering model assigns each postal code to one of five groups.



**Results:**



The results of the clustering analysis are displayed on a map to visualize the postal regions. Out of the five clusters, two only include a single postal code. The cluster with a single postal code has a unique venue that defines the postal region such as postal code 64195 containing the Kansas City International Airport. Both single postal code clusters are set aside so the focus can be on the remaining three. The cluster displayed by red dots on the map is assigned the label “Cluster 1” and is defined by the different types of bars and entertainment venues in the area. Cluster 1 is named “City Centre” because each postal region is in a high-density area that caters to nightlife and tourism. The next cluster marked on the map by blue markers is labeled “Cluster 2.” Cluster 2 is called Industrial Oldtown because it is defined by antique shops, border crossings, and parks that cater to the old town and blue-collar community. The green markers labeled “Cluster 3” or “Suburbs” is defined by an abundance or restaurants and shopping centers. The Suburb cluster caters to residents who value space and quiet over community and entertainment.

**Discussion and Conclusion:**

The cluster I labeled "Industrial Oldtown" seems best suited for a community coffee shop. The antiques and other local venues demonstrate a high level of community value. The area seems to support local shops over nationwide chains more than the average community. The cluster contains a constant flow of potential customers because an abundance of industrial jobs attracts workers who live outside of the area. More specifically, the coffee shop should look to open within the postal code 64109 region because it remains the only area without a coffee shop as a top 10 common venue. Postal code 64109 makes up the southern portion of an area known as the West Bottoms. The combination of local support and less competition make the southern portion of the West Bottoms community in Kansas City an ideal place to open a community and relationship driven coffee shop.