## horizontal line



Moving from Indore to Raleigh­­: Relocating with Data Science

09.07.2020

**─**

Anmol Lunavat

IBM Applied Data Science Capstone

1. **Introduction**

Each year there are a number of people who move from one place to another for variety of reasons such as a new job, education, or retirement. Moving to a new city though exciting comes with many challenges. Since we are accustomed to the place and neighborhood we have been staying in, it is easier to make this shift if the new city has a familiar environment. There are multiple factors which constitute in deciding whether two cities are similar such as food, transportation, climate, culture etc. Though, factors such as climate or culture aren’t under our control, we can analyze our new city based on food preferences and venues. Therefore, it will be beneficial to predict similar neighborhoods between two cities based on their venues.

Venues surrounding a neighborhood can be analyzed on location and category. Additionally, venue categories can be used to classify similar neighborhoods together. The aim of this project is to predict neighborhoods in Raleigh which are similar to neighborhoods Vijay Nagar and Old Palasia in Indore. The project implements a clustering method which utilizes different categories of venues to group neighborhoods in Raleigh.

The findings of this project will interest people planning to move from Indore to the city of Raleigh.

1. **Data Collection and Strategy**

To analyze the similarity between neighborhoods of Raleigh with Vijay Nagar and Old Palasia in Indore we gathered a list of neighborhoods and venues in Raleigh.

* 1. Neighbourhoods

A list of neighborhoods in Raleigh could be found on the Wikipedia page Raleigh, North Carolina Neighborhoods. Python web scraping techniques with URL handling and beautifulsoup packages was utilized to extract the list and convert it into a pandas dataframe. Following this, Python geocoder package was used to collect latitude and longitude values for all the above collected neighborhoods.

* 1. Venues

In the second part of data collection we utilized Foursquare API to gather a list of venues in the neighborhoods of Raleigh. Particularly we relied on Places by Foursquare, a database of more than 105 million places worldwide and API services that enable retrieval of location data. Format of the data received was not appropriate for the purpose of analytics. Thus, we employed data wrangling techniques to transform the data into a more appropriate and valuable format.

1. **Methodology**

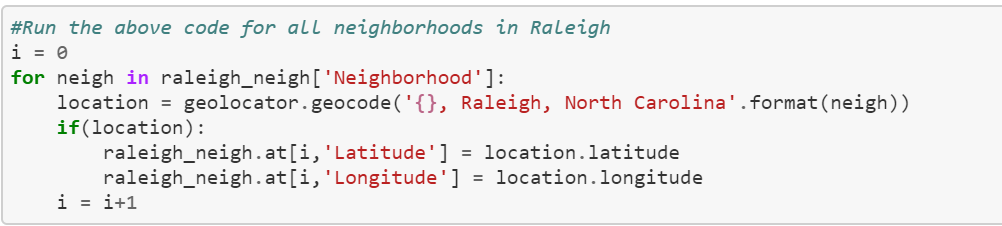
In order to find similar neighborhoods in Raleigh we were required to analyze the venues in each neighborhood and cluster them accordingly. Following this, we were able to use the data of venues in Vijay Nagar and Old Palasia to predict which cluster of neighborhoods they belong to.

* 1. Neighborhoods in Raleigh

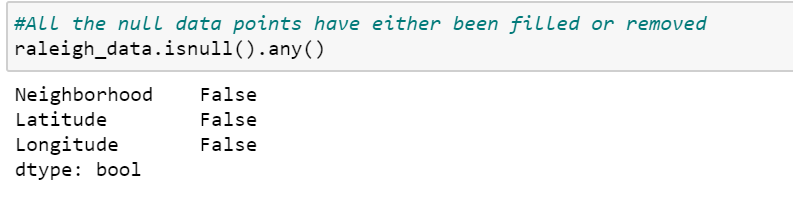
After collecting the data from Wikipedia, we converted it into a Pandas dataframe. There is a total of 105 neighborhoods in the City of Raleigh.

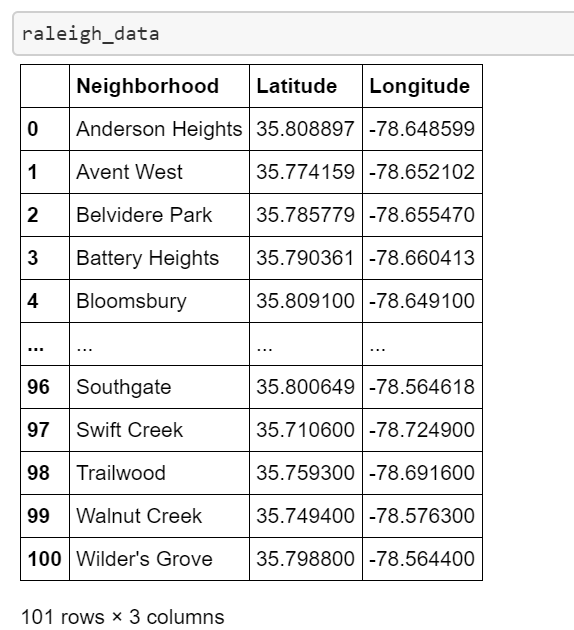


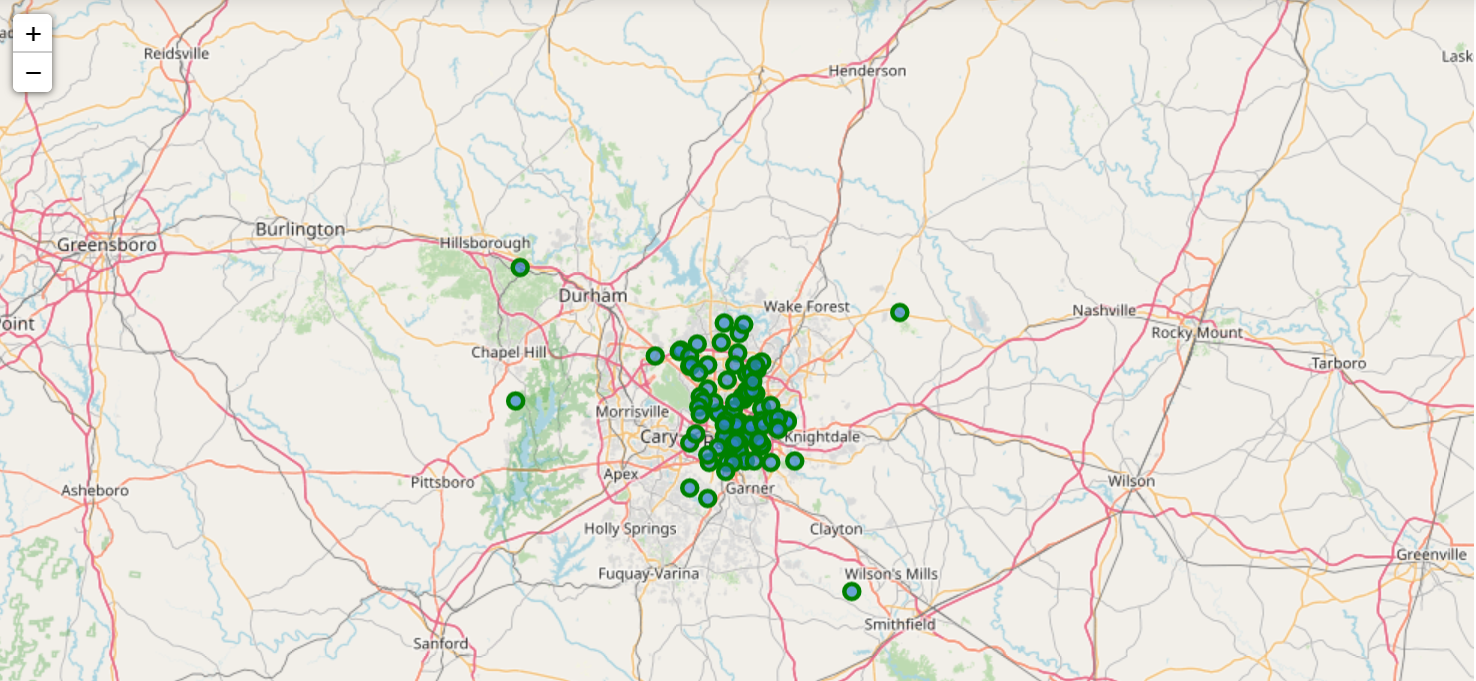
To utilize Foursqaure API and collect a list of venues we would require longitudes and latitudes of the all the neighborhoods. We utilized Geolocator geocode to fetch the required data for each neighborhood.



Upon analyzing the accuracy of the Geolocator we found out 15 neighborhoods which did not have a latitude or longitude assigned to them. For the state of accuracy, we filled the missing values by looking them up by ourselves and skipped the ones for which the data was not found.







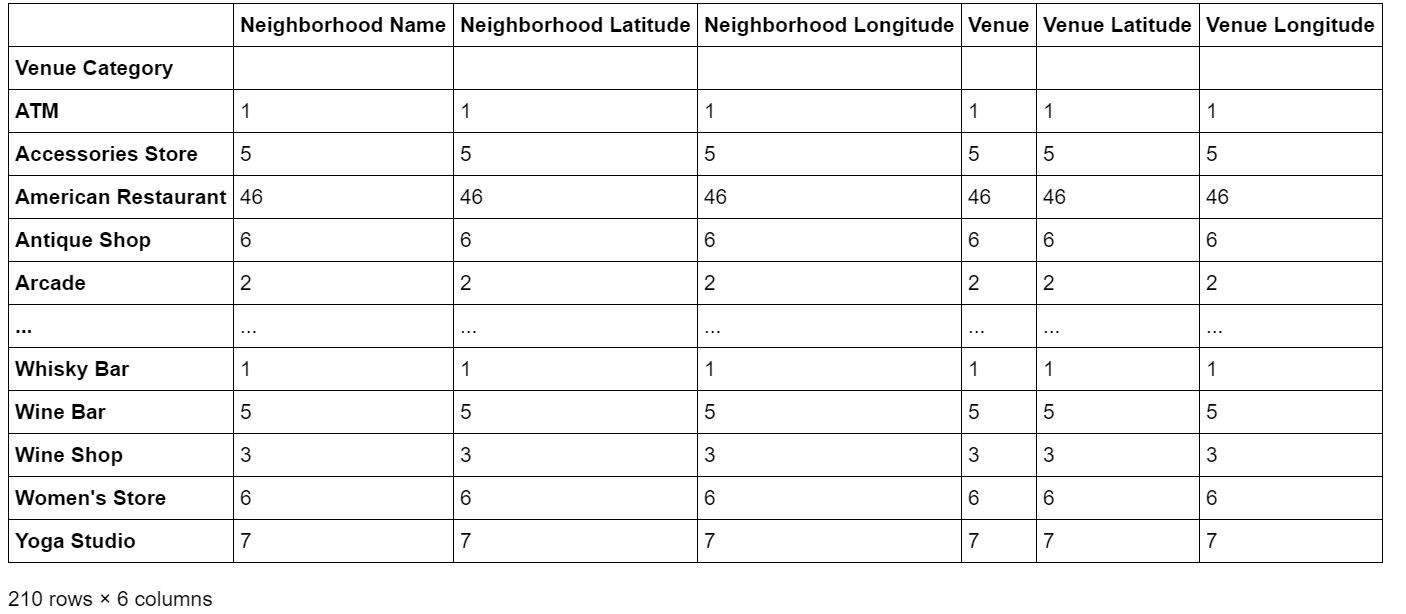
Upon completion of the above steps our final dataset had a total of 101 neighborhoods from the city of Raleigh.

* 1. Venues in the Neighborhoods of Raleigh

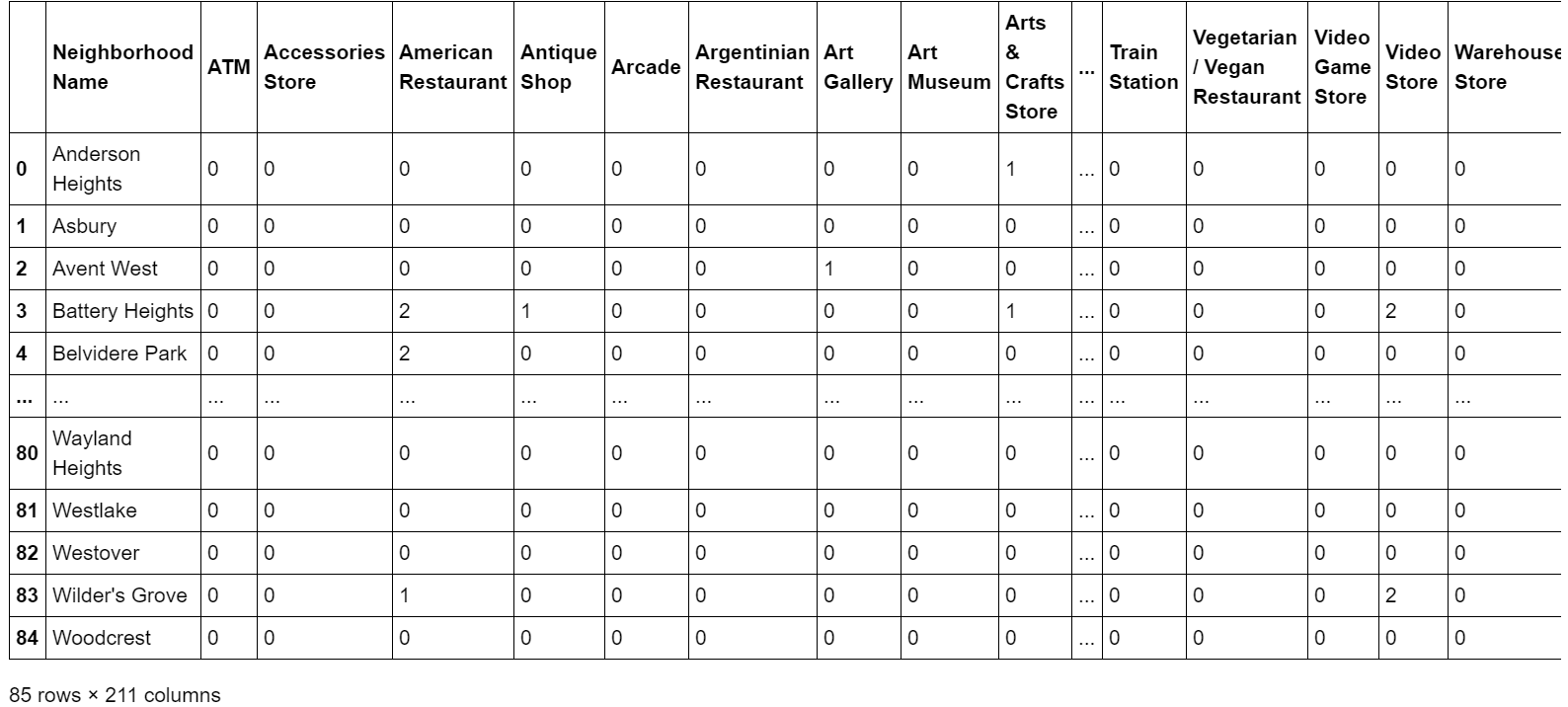
To fetch a list of venues in the neighborhoods of Raleigh we utilized Foursqare API. From the result set we collected Venue Name, Category, Latitude and Longitude of the venue.



Venue category was an important aspect of the analysis. We used this particular feature to cluster the neighborhoods and find similarities between Raleigh and Indore. While analyzing the result set, we found out that we have a total of 210 unique venue categories in Raleigh.

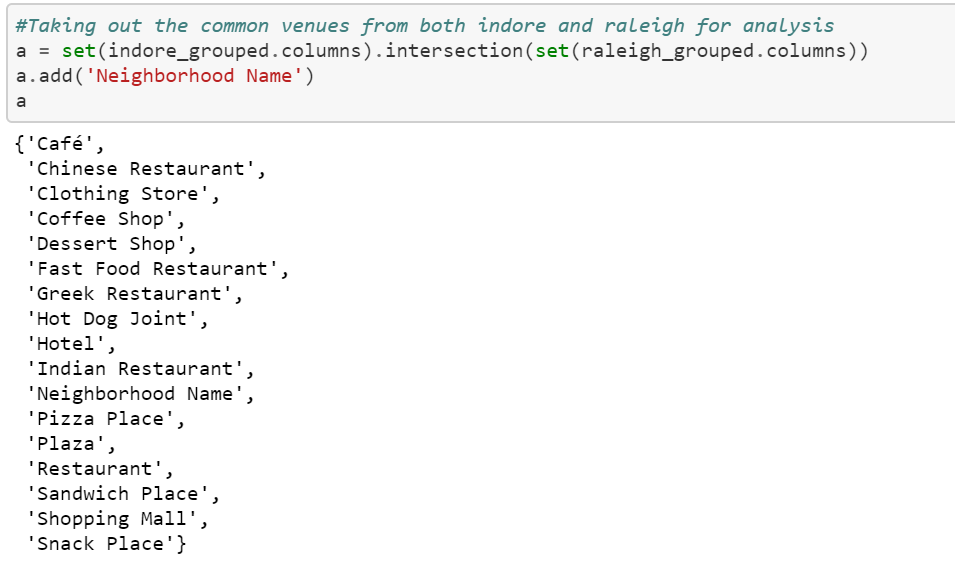


Following this, we converted the data into one hot encoded set and grouped them on the basis of neighborhood. While implementing grouping, we took the sum of each category in a particular neighborhood respectively.

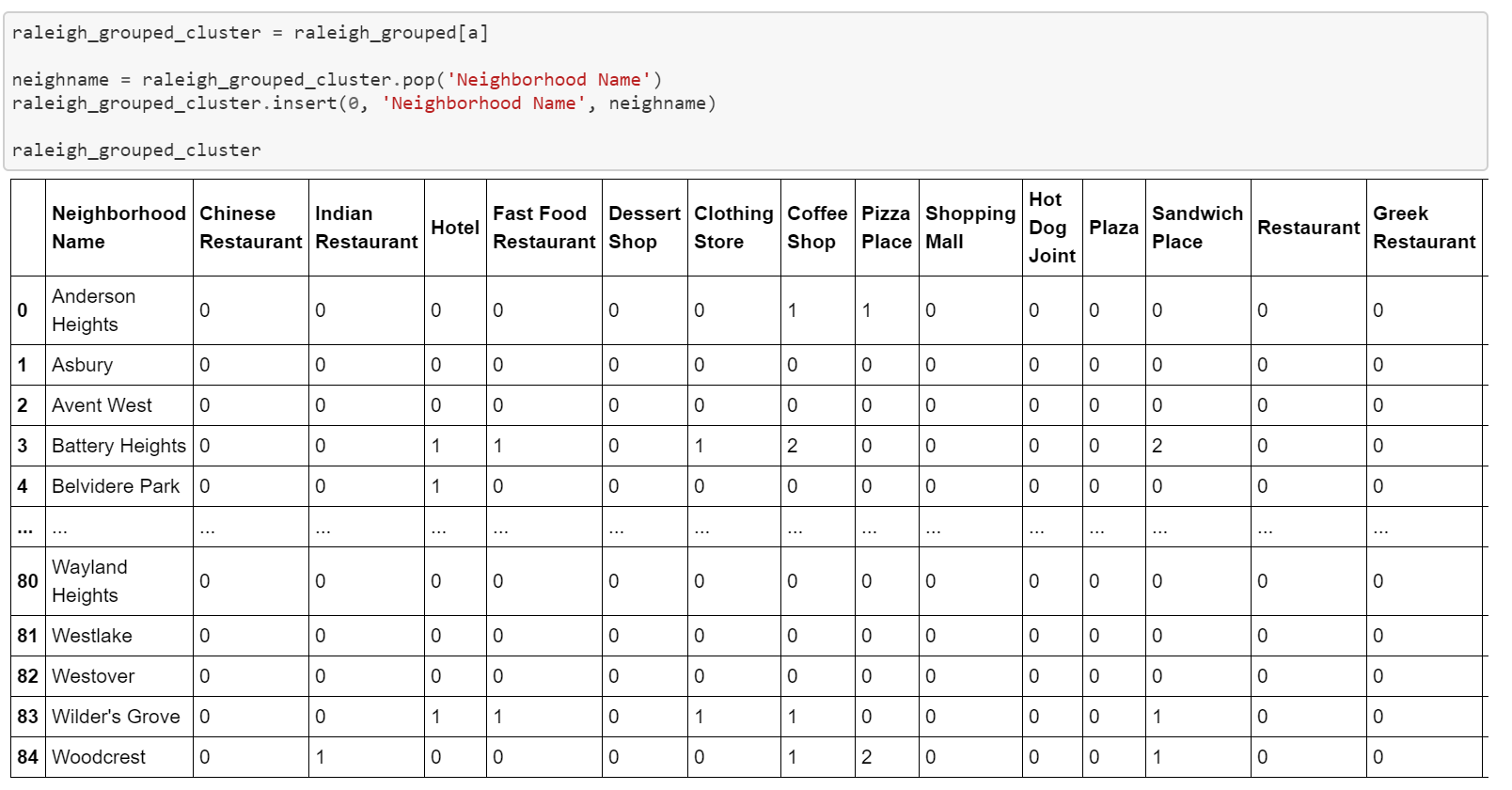


* 1. Common Venue Categories in Raleigh and Indore

Following a similar process, we fetched the venues in Vijay Nagar and Old Palasia. To create a common clustering mechanism, we analyzed the common venue categories in both data sets. This was done to ensure that the algorithm trained on Raleigh is also able to predict cluster labels for Vijay Nagar and Old Palasia.



Since we wanted to predict cluster labels for Indore, it was important to ensure we were considering only the venue categories which are common between both. Otherwise, an algorithm trained on actual 210 venue categories of Raleigh would have not been able to predict cluster labels for Indore as it had lesser number of categories. For our model to work accurately we require both our data sets to have the same number of features. Therefore, the next step was to remove uncommon venue categories from the data set of Raleigh before we could run our clustering algorithm on this data.

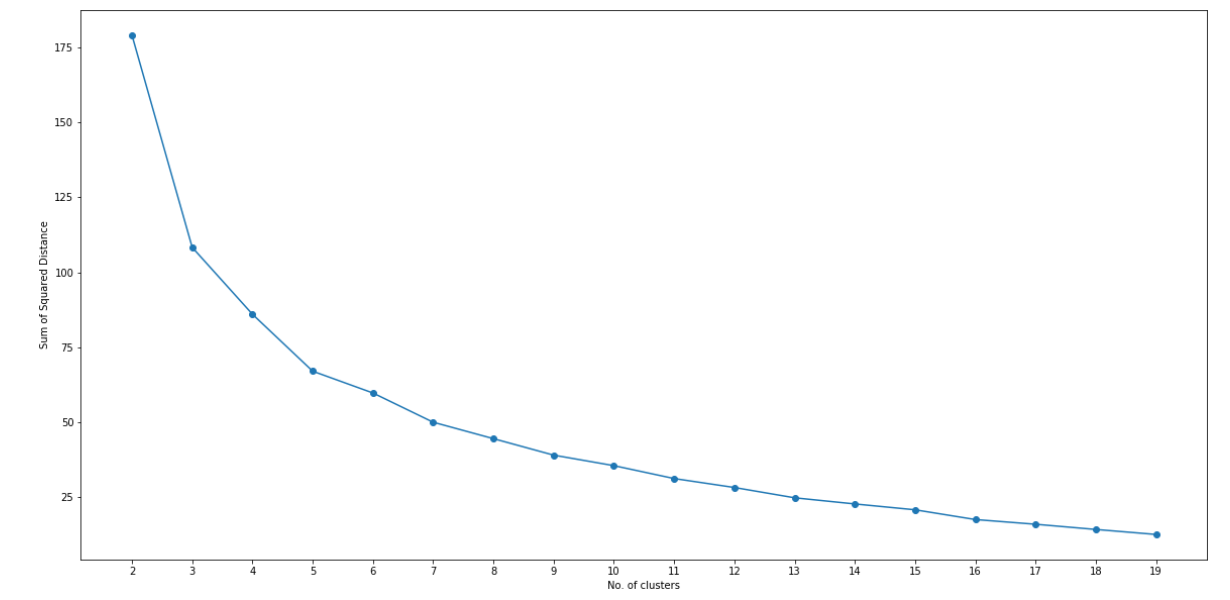


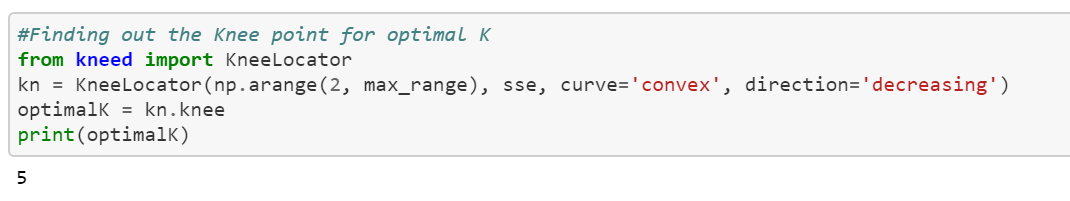


1. **Modelling**

K-means Clustering algorithm was utilized to cluster the neighborhoods of Raleigh. Clustering the data identified subgroups in the data set such that neighborhoods in the same cluster were very similar while neighborhoods in different clusters were very different.  Clustering requires a feature which could be used to find subgroups amongst the data. In our use case, the feature venue categories was used to find best subgroups of neighborhoods. Once we found the common value for the feature venue categories between Indore and Raleigh, we were able to fit the algorithm and cluster neighborhoods of Raleigh.

* 1. Optimal number of K

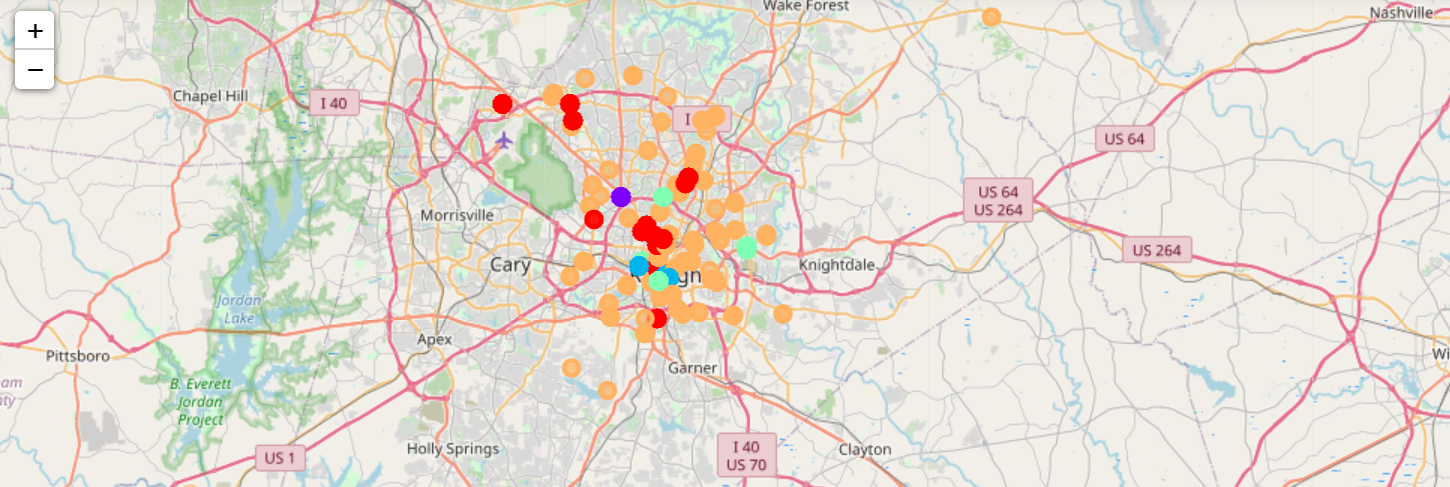
To find the optimal number of clusters for our data we calculated Sum of Squared Distance. We ran for K in range of 2 to 20 and drew an elbow curve to find the best value of K.



As visualized the knee of the curve occurs when the number of clusters is equal to 5. Thus, we used 5 as the value of K for our algorithm.

* 1. K-means Clustering

We ran K-means on our data and clustered the neighborhoods in Raleigh into 5 different clusters. Each cluster contained neighborhoods which were similar to each other based on the categories of venue it had.









1. **Results**

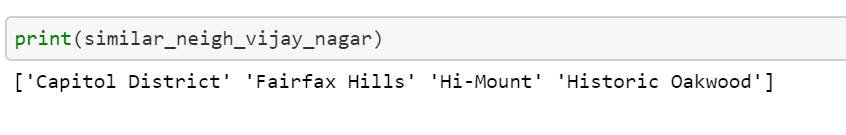
We found that most of the neighborhoods either belonged to Cluster 0 or 4. There were 65 neighborhoods in Cluster 0 and 10 in Cluster 4. After our kmeans algorithm was trained on neighborhoods of Raleigh, we utilized it to predict cluster label for Vijay Nagar and Old Palasia to find their similarity with neighborhoods of Raleigh.



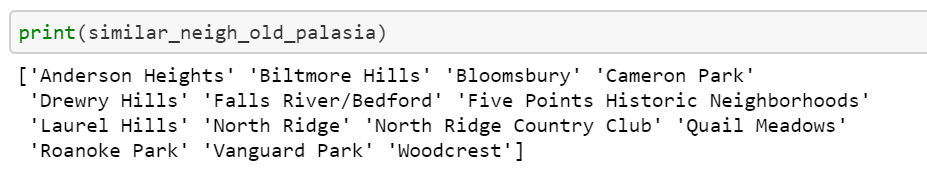
While Old Palasia belonged to cluster 2, Vijay Nagar belonged to cluster 0.

Thus, we were able to find neighborhoods similar to both Vijay Nagar and Old Palasia by analyzing the other members of their cluster.

A person residing in Vijay Nagar, Indore will find the similarity in the following neighborhoods if he visits Raleigh:



Whereas, a person residing in Old Palasia, Indore will find the similarity in the following neighborhoods if he visits Raleigh:



1. **Future Work**

The current scope of the project only considers two locations from Indore, i.e., Vijay Nagar and Old Palasia. The inclusion of all neighborhoods from Indore will make us capable of comparing two cities and suggesting similar neighborhoods for each neighborhood from Indore. Furthermore, the system could become capable of running between any two or more cities and suggesting similar neighborhoods between them. Once the system is capable of dynamically adapting to multiple cities it could be used as a platform for finding similar neighborhoods. A user would be allowed to enter current city and destination cities. The system would then utilize the current city and suggest similar neighborhoods in the destination cities.