An Analysis of Boston, MA Neighborhood Restaurant Data

IBM DATA SCIENCE CAPSTONE PROJECT

Shaun Anderson

Intro/Business Problem

- The city of Boston, Massachusetts is a small city on the northeast coast of the United States. The city is well known for its many cultural and historical attractions that draw in approximately 20 million visitors per year from both the United States and abroad
- The city's tourist attractions tend to be contained in one small area of the city with more residential neighborhoods around that. The tourist areas do have some residential areas but these tend to be very upscale and expensive. The residential areas outside of that zone tend to be more affordable and are where middle-class and college students tend to reside.
- Analysis of these neighborhoods will be performed to see if there is a difference between the restaurant types that are most prevalent in the tourist areas versus the residential areas. This would assist someone who wishes to open a restaurant in these areas in determining which type of restaurant may be successful.

Data Acquisition

- Zip code (postal code) and neighborhood data for the city of Boston was scraped from the web page: https://en.wikipedia.org/wiki/Boston#Demographic_breakdown_by_ZIP_Code.
- The geopy Nominatim() function was used to add latitude and longitude coordinates based on the neighborhood's zip code.
- Foursquare API was called to gather restaurant data for the neighborhoods

Analysis

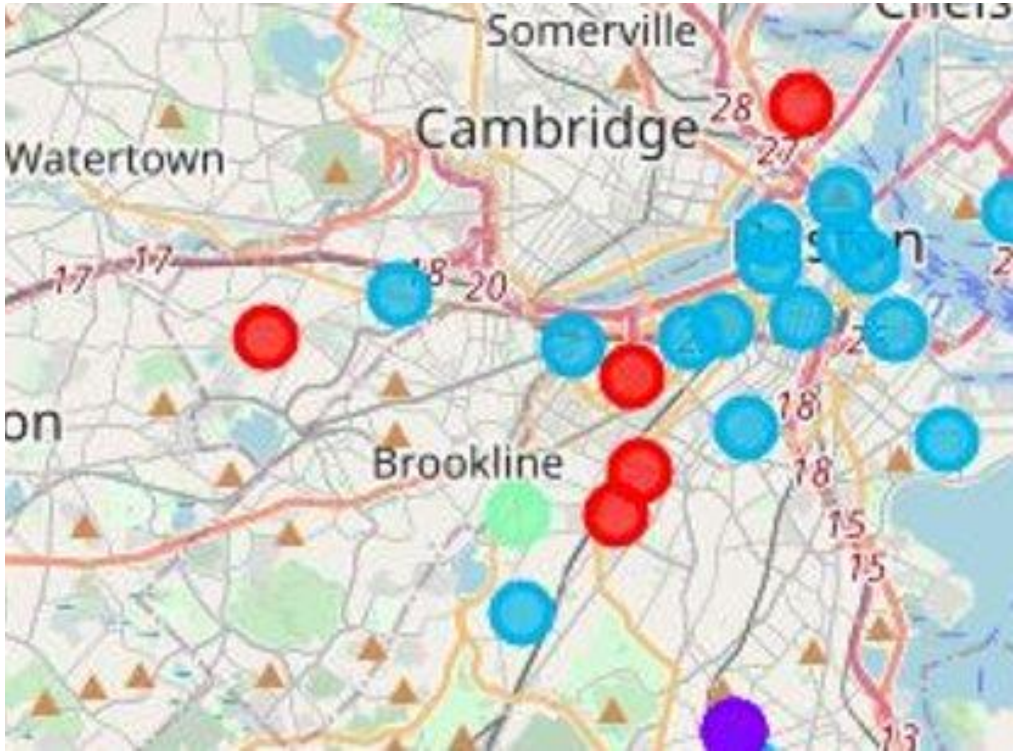
One-hot encoding, groupby() and mean() was used to get the frequency of the different restaurant categories in the neighborhoods

The top 5 restaurant categories in the neighborhoods was determined

Cluster analysis using the k-means unsupervised machine learning cluster algorithm was performed using values of k=5 and k=7

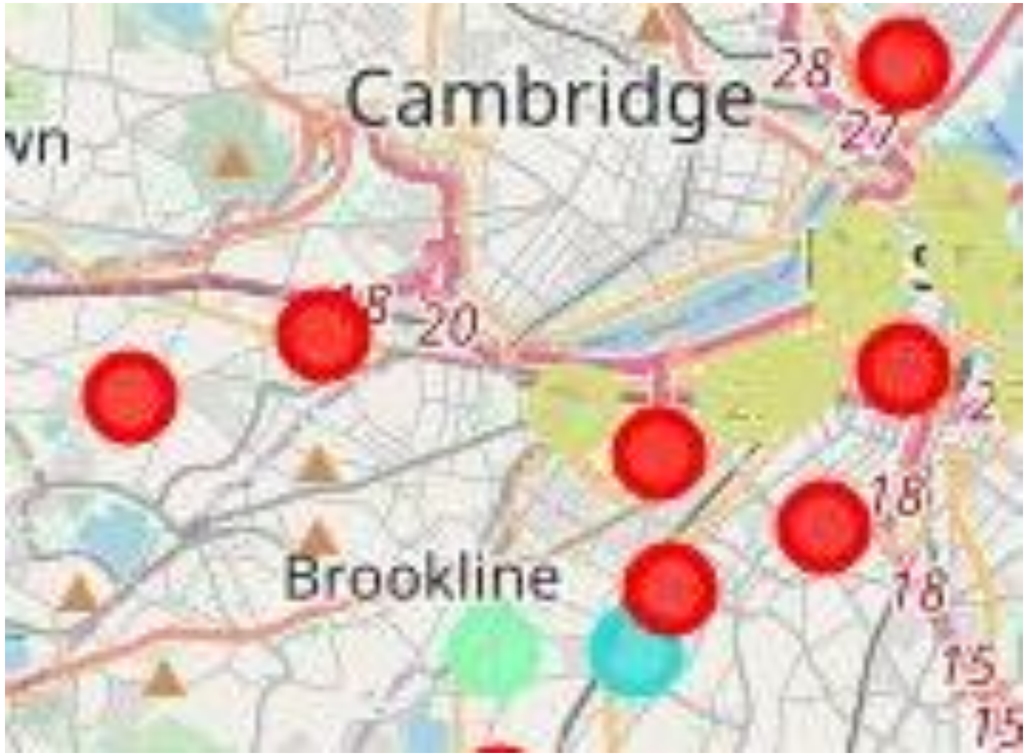
Analysis k=5

- Using k=5 resulted in no real delineation of the neighborhood restaurant types



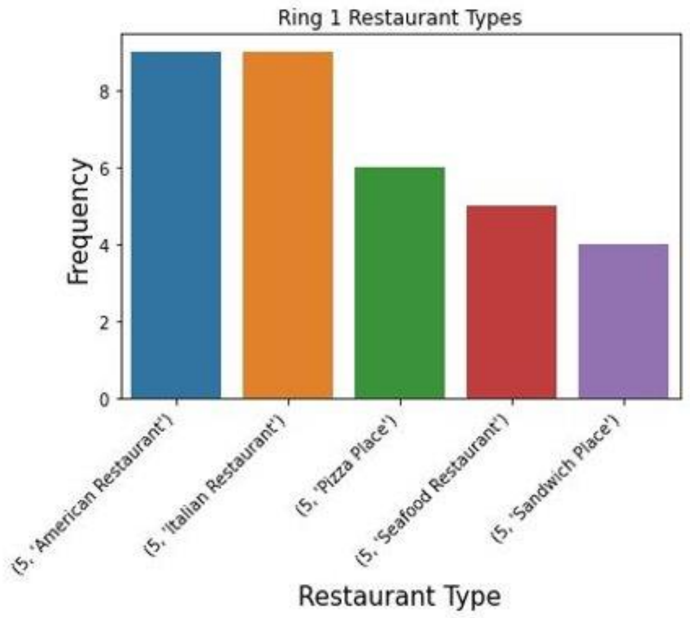
Analysis k=7

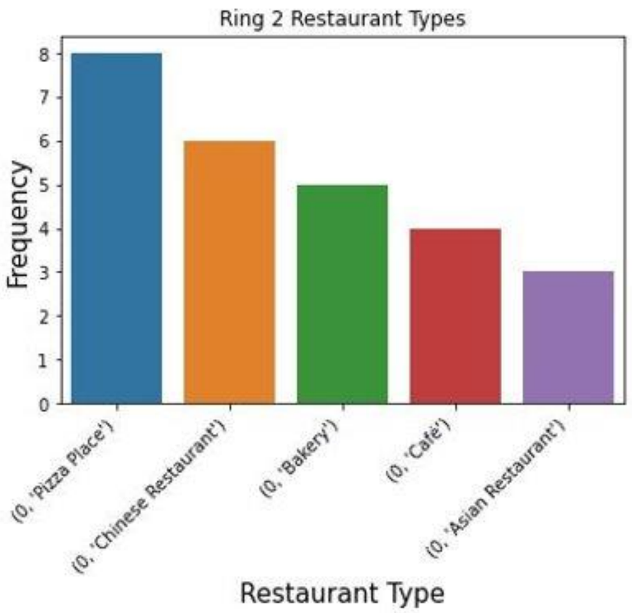
Using k=7 resulted in a clear delineation of the neighborhood restaurant types. The tourist area is gold, the residential is red



Neighborhood Comparison

- The two neighborhoods have quite different restaurant types:
- The top graph, tourist areas and upscale residential, have more upscale, sit down style restaurants
- The lower graph, residential areas, have more low cost and take-out options





Conclusion

- Using my knowledge of the city as a reference these results appear to make sense. In the previously shown maps below, the gold colored dots are the tourist areas whereas the red dots are the residential neighborhoods.
- As shown in the bar charts above the restaurant types in the tourist (and upscale residential) area tend to be more upscale, expensive, sit-down type establishments such as American, Italian and seafood.
- The residential areas appear to be more lower priced categories and seemingly would have take-out options such as pizza and Chinese. Also notable are more bakeries and cafes where college students would study and commuters would stop on the way to work to grab a coffee.
- A person or company wishing to open a restaurant in these areas could use this information to determine the type they should open in order to be most successful. For example, an upscale seafood restaurant may not be successful near apartments occupied by college students.

Discussion

- This basic analysis seems very accurate in its ability to cluster these 2 areas correctly.
- The city of Boston has irregular city limits which I believe may have hindered the analysis somewhat. My other clusters tended to be more isolated.
- I believe that including the cities of Cambridge and Brookline could help the model cluster areas more efficiently as they would provide a more regular-shaped area for analysis.
- At the same time, Brookline is a very expensive area wedged between 2 of Boston’s residential areas and might skew the model even worse.
- Additionally, population, demographic and income data could be used to expand the model input and possibly provide a more detailed conclusion.