Applied Data Science — Clustering Neighborhoods in Staten Island, NY

Chris Calamera – 1/29/2020

Introduction/Business Problem

Staten Island is a borough of New York, located on the southern tip of Manhattan. It houses about 750,000 people across roughly 8 ½ x 14 miles of land.

While growing up on Staten Island, my cover band used to enjoy playing at various venues across the area. Since those days, the music scene on Staten Island has all but dried up. With my project, I am going to attempt to find where it would be possible to open a music venue on Staten Island that would cater to the types of crowds that enjoy coffee and similar amenities – thus, the inception of a potential "Coffeehouse Music Venue".

Methodology

- **Data Collection**

There is no available dataset in csv/json/xls format which contains all of the with latitude and longitude data for the neighborhoods of Staten Island. This data is only available in Wikipedia, and my project will use a web scraper to collect the data from the Wikipedia page for Staten Island neighborhoods and populate the data appropriately for analysis.

Example page:

https://commons.wikimedia.org/wiki/Category:Neighborhoods_in_Staten_Island,_New_York_City

Extracting Coordinates from Web scrapped Data

We will now start prepping the data from the Staten Island Wiki page. We will use our python code to organize the data into a usable format – with this, we see that there are 66 unique neighborhoods in Staten Island.

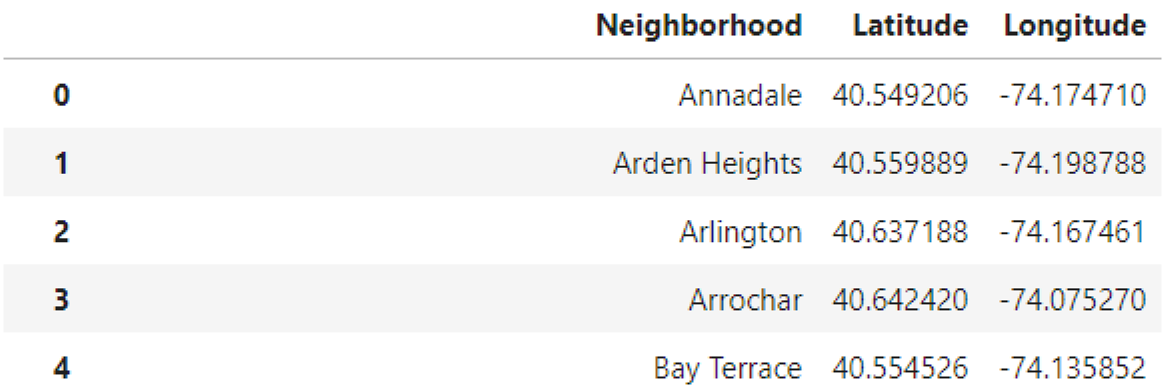
Issues faced during Extraction:

1. Extracted data doubled the data return on the neighborhoods, so only the first 66 rows needed to be kept, with the rest being deleted.

2. Extracted data featured a few records that were scraped from some random spots on the page – these were deleted.

Extracted Data frame

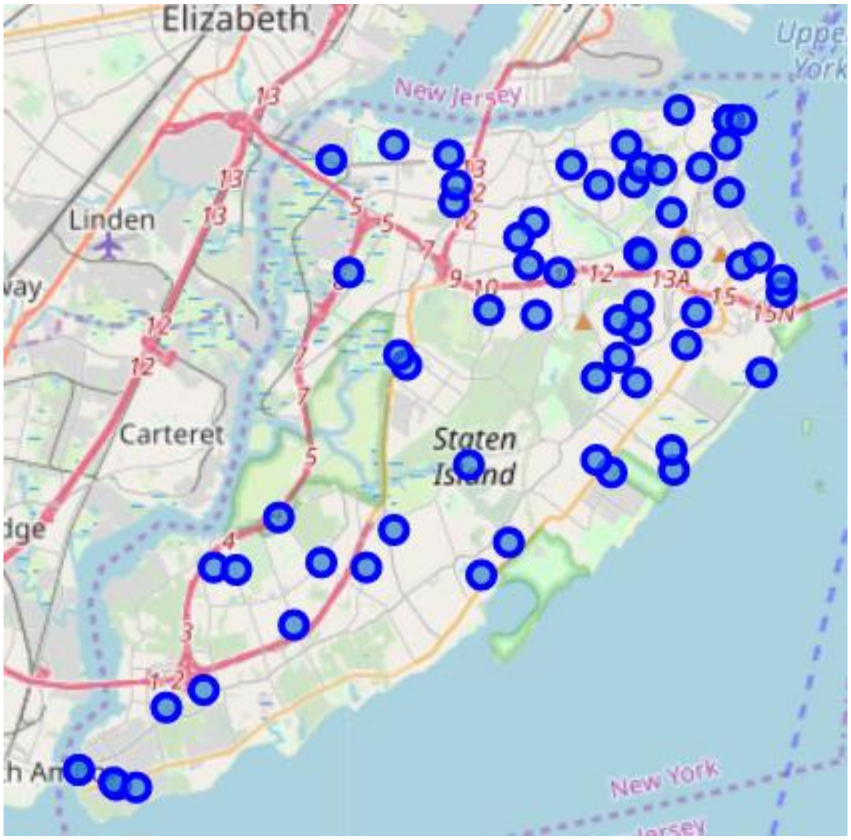
The below output will show how the extracted data will look in a data frame:



	Neighborhood	Latitude	Longitude
0	Annadale	40.549206	-74.174710
1	Arden Heights	40.559889	-74.198788
2	Arlington	40.637188	-74.167461
3	Arrochar	40.642420	-74.075270
4	Bay Terrace	40.554526	-74.135852

Projection of Data Points on Folium Map

Here we see a map of Staten Island with blue circle, representing each neighborhood in Staten Island. The "Folium CircleMaker" method has been used to create these circles by passing latitude and longitude data from our web scraped data frame. The "Folium Popup" method has been used to create pop-ups with each neighborhood name. This popup will be visible whenever a user clicks on these blue circles.

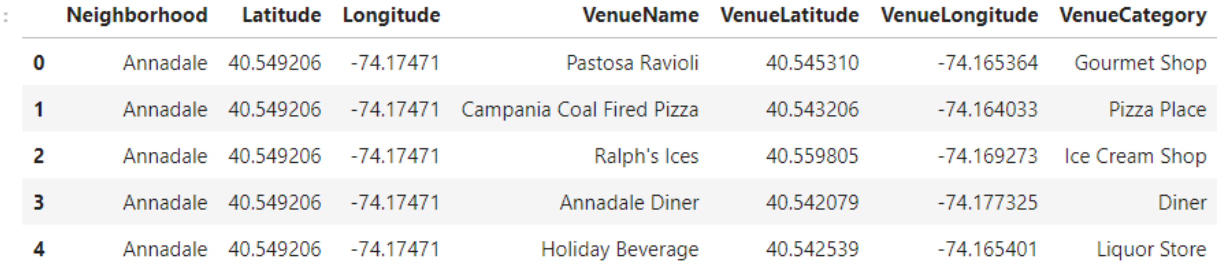


Example Data Frame

This is how an example data frame may look from the Foursquare data pull. There are numerous fields returned, but we will only considered certain ones for our project.

After analyzing Foursquare Explore API data set (which can be accessed with the below API call):

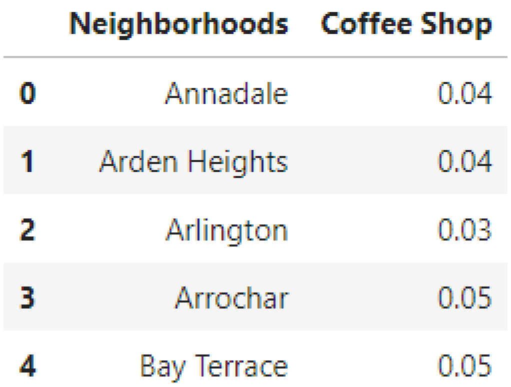
https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}



	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Annadale	40.549206	-74.17471	Pastosa Ravioli	40.545310	-74.165364	Gourmet Shop
1	Annadale	40.549206	-74.17471	Campania Coal Fired Pizza	40.543206	-74.164033	Pizza Place
2	Annadale	40.549206	-74.17471	Ralph's Ices	40.559805	-74.169273	Ice Cream Shop
3	Annadale	40.549206	-74.17471	Annadale Diner	40.542079	-74.177325	Diner
4	Annadale	40.549206	-74.17471	Holiday Beverage	40.542539	-74.165401	Liquor Store

After collecting nearby venues for available geographical points in Staten Island, those points will then be filtered to consider only categories which can be marked as “Coffee Shops”.

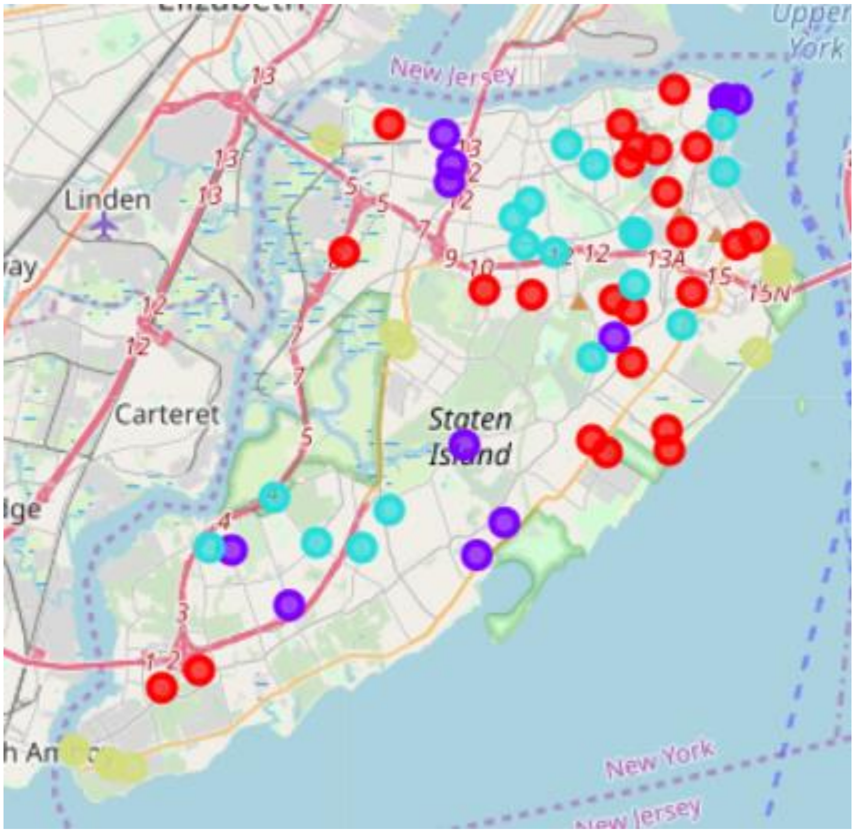
Example Data Frame – Coffee Shops only



	Neighborhoods	Coffee Shop
0	Annadale	0.04
1	Arden Heights	0.04
2	Arlington	0.03
3	Arrochar	0.05
4	Bay Terrace	0.05

K-Means Clustering

K-Means clustering is one of the most common cluster methods of unsupervised learning. In our project, we will run K-Means to cluster the neighborhoods into 4 clusters. You can see the clusters highlighted in the map below:



Conclusion:

What we were trying to do in this exercise is to find areas on Staten Island with coffee shops in a nearby cluster.

This gives us an opportunity to explore opening a coffeehouse environment for music and using some of the nearby vendors as collateral for boosting/driving business.¶

Areas near Silver Lake and Meier's Corners represent potential landing spots for this business venture.