Applied Data Science Capstone Project Report

-John Linskey

1. Introduction

This project is aimed to provide a local bicycle business crucial information in their decision making concerning opening a new retail location within the Atlanta Metro area. Specifically, the business owner is requesting detail information comparing the two cities of Smyrna and Cedartown, due to their proximity near to or connecting to the Silver Comet trail.

1.1 Background

With the recent closure of all Performance Bicycle retail stores (the largest national retailer of bicycles and accessories in America), the Atlanta metro has lost 4 major bicycles shops. Today, many people within the metro community are without any bicycle shop due to the closures.

The Silver Comet and Chief Ladiga trails join to form one continuous 94-mile (151 km) trail from Smyrna, Georgia (Atlanta area) to Anniston, Alabama, which together form the second longest paved rail-to-trail in the U.S. The trail is heavily used throughout the year by cyclist, joggers/runners, and walkers. The population along the trail varies with the closeness to Atlanta, with the nearest trail heads being more widely used. The western (Georgia) sections of the trail are less densely populated, and hopefully this report will concluded that there are opportunities in either neighborhood.

1.2 Problem

A local bicycle shop in the south metro community of Peachtree City, Hometown Bikes, is investigating into expand into the Atlanta Metro area. They see that the time is right to expand their business. In addition to expanding, they are aiming to be located near the Silver Comet trail in order to provide support and services to travelers along the trail. In addition to providing bicycle service and parts, the business is considering providing other such services as coffee, food, guided rides (both weekly and monthly) and into future overnight accommodations. They are requesting location information of competitors' venues within the two cities as well as the popularity of the venues.

Hometown Bikes is only considering locations within the state of Georgia, therefore they have picked Cedartown as the western most potential location.

2. Data acquisition and cleaning

2.1 Data sources

I started by obtaining the geo coordinates for both the cities of Smyrna and Cedartown from Wikipedia.

In addition, I wanted to map a few of the trail heads along the trail for a reference guide. The data was easily obtained from the following source.

```
#List of trail heads #http://wikimapia.org/#lang=en&lat=33.893787&lon=-84.737549&z=12&m=w&show=/street/15589747/Silver-Comet-Bicycle-Trail
```

I then made use of the Foursquare Venue Search API for both cities to obtain venue information. Due to the differences in population density, the search radius of Smyrna was 5 miles; whereas the search radius in Cedartown was 20 miles.

Afterwards, I used panda dataframes to conduct clustering to identify each area and its category. The data source that I made use of is from Foursquare, which is a technology company that uses location intelligence to build meaningful consumer experiences and business solutions.

The business owner was interested in reviewing the venue location information for the following Foursquare categories:

Bike shops: 4bf58dd8d48988d115951735

Sporting Goods Shop: 4bf58dd8d48988d1f2941735

Coffee Shops: 4bf58dd8d48988d1e0931735
Campgrounds: 4bf58dd8d48988d1e4941735
Hotels: 4bf58dd8d48988d1fa931735

Bus Station: 4bf58dd8d48988d1fe931735

These can be found on the Foursquare Developer resources page: https://developer.foursquare.com/docs/resources/categories

2.2 Data cleaning

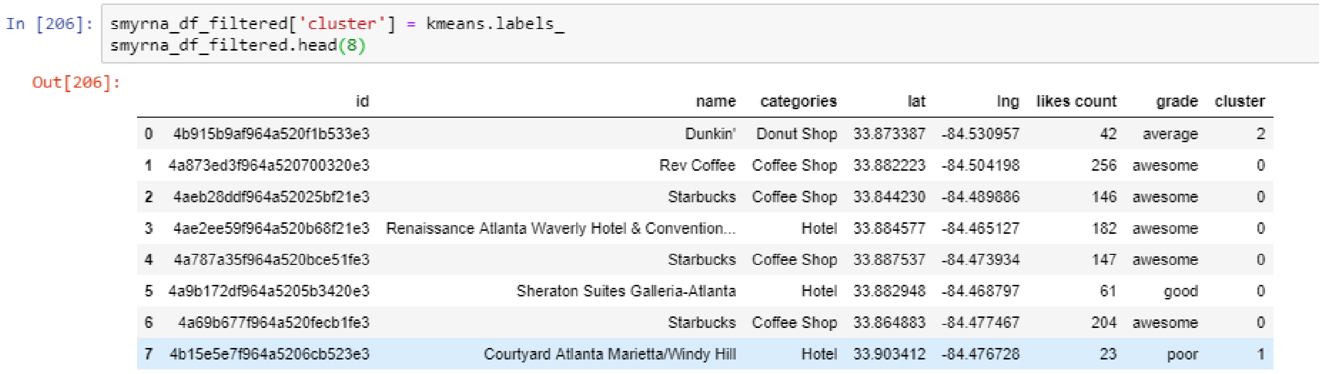
After obtaining the GeoJson files, I prepared the data sets by cleaning and transforming. The dataframes were then created for the data sets.

2.3 Feature selection

For the initial steps, the primary columns that I was concerned with included the 'id', 'name', 'categories', 'location.latitude', and 'location.longitude'.

3. Exploratory Data Analysis 3.1 Calculation of target variable

Addition data was retrieved from Foursquare, so that I could assign ‘likes count’ to each of the venues. After assigning the counts, I conducted grading on each of the venues, which later is used in the clustering.



In [206]:

```
smyrna_df_filtered['cluster'] = kmeans.labels_
smyrna_df_filtered.head(8)
```

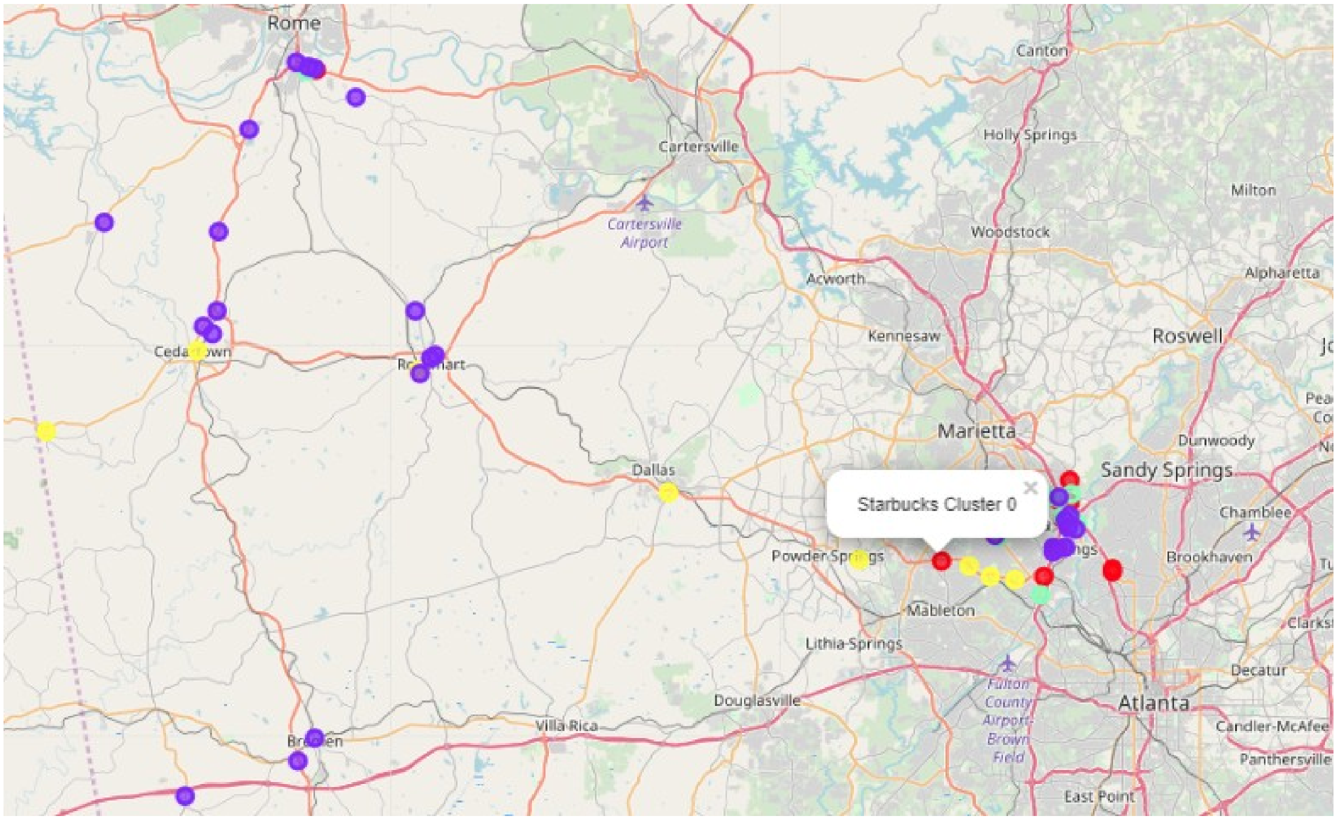
Out[206]:

	id	name	categories	lat	lng	likes count	grade	cluster
0	4b915b9af964a520f1b533e3	Dunkin'	Donut Shop	33.873387	-84.530957	42	average	2
1	4a873ed3f964a520700320e3	Rev Coffee	Coffee Shop	33.882223	-84.504198	256	awesome	0
2	4aeb28ddf964a52025bf21e3	Starbucks	Coffee Shop	33.844230	-84.489886	146	awesome	0
3	4ae2ee59f964a520b68f21e3	Renaissance Atlanta Waverly Hotel & Convention...	Hotel	33.884577	-84.465127	182	awesome	0
4	4a787a35f964a520bce51fe3	Starbucks	Coffee Shop	33.887537	-84.473934	147	awesome	0
5	4a9b172df964a5205b3420e3	Sheraton Suites Galleria-Atlanta	Hotel	33.882948	-84.468797	61	good	0
6	4a69b677f964a520fecb1fe3	Starbucks	Coffee Shop	33.864883	-84.477467	204	awesome	0
7	4b15e5e7f964a5206cb523e3	Courtyard Atlanta Marietta/Windy Hill	Hotel	33.903412	-84.476728	23	poor	1

4. Predictive Modeling

Classification models

Using techniques such as K-means clustering, I will able to get results about competitors’ venues within the two areas that would be attractive to the business owner. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.



5. Conclusions

By making use of Foursquare data sets and data science methodology, I have successfully created information to assist the business owner in selecting a new retail location that will serve the needs of travelers along the Silver Comet Trail.

From my research, it appears that both areas are in need of a bicycle shop, because I didn’t located any currently opened bike shops.

Further study concerning the population of the areas, see U.S. Census Bureau QuickFacts, indicates that Hometown Bike’s current location has roughly a population of 35,766 people. The 20 mile radius around Cedartown has roughly the population of 46,849 (Rome + Cedartown); whereas Smyrna has a population of around 56,706.

Based upon the research data of the location venues, clustering, and Census population data, I would conclude that Cedartown might be a better location opportunity for Hometown Bikes. This area has nearly the same amount of population, and currently has competitors' venues. In addition being that this is the western most trail head to the Silver Comet Bicycle Trail, this location very well could serve as a stay overnight spot for travelers attempting to ride the entire trail. In addition, Hometown Bikes would have other opportunities to provided other services such as snack/coffee bar, and guided rides.