IBM Coursera Capstone Project

In Data Science

Capstone project final report

Atlanta city neighborhood and its surrounding venues Analysis

By

Ravindra Bhagwat – Dec 2019

Table of contents

I. Introduction: 2
II. Data description: 3
III. Methodology: 6
1. **Visualization using Folium**: 6
2. **KNN clustering and segmentation**: 7
IV. Results: 9
V. Discussion: 10
VI. Conclusion: 11
References: 12

I. Introduction:

This report is for the IBM Capstone project final course of the Data Science Specialization. A 9-courses series created by IBM, hosted on Coursera platform. The problem and the analysis approach are left for the learner to decide, with a requirement of leveraging the Foursquare location data to explore or compare neighborhoods or cities of your choice or to come up with a problem that you can use the Foursquare location data to solve.

The main goal will be exploring the neighborhoods of Atlanta and vicinity cites in order to extract the type and density of venues in Atlanta and its surrounding neighborhoods.

This analysis will help a small-scale business entrepreneur to start a new business in these neighborhoods or for a family to find a place with the amenities that they need to enjoy the stay after moving to Atlanta. It's common that the owners or agents advertise their properties are closed to some kinds of venues like supermarkets, restaurants or coffee shops, etc.; showing the "convenience" of the location in order to raise their house's value or business value.

So, can the surrounding venues affect the profitability of the business and become successful. So, what types of venues have the most affect, both positively and negatively need to be analyzed further.

The target audience for this report are:

- Small and medium business entrepreneurs who want to start a related business that can be successful in the neighborhood.
- Potential buyers who can roughly estimate the best value for a house they are buying based on the surrounding venues.

II. Data description:

Atlanta city neighborhoods were chosen as the observation target due to the following reasons:

- The availability of business venues and prices vary and are limited in some areas.
- The diversity of venues and prices varies between neighborhoods. For example, a business in downtown Atlanta may cost lot more on average; while in Sandy springs, Alpharetta, Johns Creek, just 45 minutes away, it will be lot cheaper and has many potential customers.
- The availability of geo data which can be used to visualize the dataset onto a map.

The dataset will be composed from the following two main sources:

- The geo json data of Atlanta area will be downloaded from GitHub site https://github.com/codeforamerica/click_that_hood/blob/master/public/data/atlanta.geojson.
- Georgia zip codes along with county and city neighborhood info is downloaded from https://www.zipcodestogo.com/Georgia/
- Neighborhood has a total of 15 counties and 37 neighborhoods. In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains the 15 counties and the neighborhoods that exist in each county as well as the latitude and longitude coordinates of each neighborhood.
- Foursquare API which provides the surrounding venues of a given coordinates.

The process of collecting and clean data:

- Download the Georgia zip codes and cities info from the website https://www.zipcodestogo.com/Georgia/
- Scraping the info directly from the website makes it difficult due to the nature of the webpage.
- Find the geographic data of the neighborhoods from the file available in GitHub, which includes the center coordinates and their border.

- Merge the zip codes and geo json data into single dataframe, so it can be processed and visualized in the notebook. The resulting dataframe looks like below (Figure 1).

[10]:

	Zip Code	Latitude	Longitude	City	County
0	30101	34.034515	-84.707349	Acworth	Cobb
1	30083	33.797412	-84.197984	Stone Mountain	Dekalb
2	30317	33.747999	-84.315586	Atlanta	Dekalb
3	30038	33.666460	-84.139855	Lithonia	Dekalb
4	30322	33.794492	-84.326570	Atlanta	Dekalb
5	30094	33.613639	-84.053557	Conyers	Rockdale
6	30314	33.757576	-84.432245	Atlanta	Fulton
7	30363	33.791004	-84.398978	Atlanta	Fulton
8	30043	33.999280	-84.009510	Lawrenceville	Gwinnett
9	30046	33.948631	-83.995764	Lawrenceville	Gwinnett
10	30071	33.939949	-84.206233	Norcross	Gwinnett

Figure 1 – Atlanta neighborhoods

- For each neighborhood, pass the obtained coordinates to Foursquare API. The “explore” endpoint will return a list of surrounding venues in a pre-defined radius.
- Count the occurrence of each venue type in a neighborhood. Then apply one hot encoding to turn each venue type into a column with their occurrence as the value.

The result dataset is a 2 dimensions data frame (Figure 2):

- Each row represents a neighborhood.
- Each column is the occurrence of a venue type.

[28]:

	Neighborhood	ATM	Accessories Store	Adult Boutique	Airport Lounge	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Arepa Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Automotive Shop	BBQ Joint	Bakery
0	Stone Mountain	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Stone Mountain	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
2	Stone Mountain	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Stone Mountain	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Stone Mountain	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	Stone Mountain	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
6	Atlanta	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	Atlanta	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	Atlanta	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	Atlanta	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	Atlanta	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 2 - Final dataset

The dataset has 50 samples and more than 300 features. The number of features may vary for different runs due to Foursquare API may returns different recommended venues at different points in time.

The number of features is much bigger than the number of samples. This will cause problem for the analysis process. Detail and counter-measurement will be discussed further in the next section.

III. Methodology:

The main assumption in starting a new business is that its value is dependent on the surrounding venue and cannot be close to a same type of venue. Thus, regression techniques will be used to analyze the dataset. The regressors will be the occurrences of venue types. And the dependent variable will be standardized average prices.

At the end, a regression model will be obtained. Along with a coefficients list which describes how each venue type may be related to the increase or decrease of a neighborhood's business value around the mean.

Python data science tools will be used to help analyze the data. Completed code can be found here: https://github.com/rdbhagwat/Coursera_Capstone/blob/master/Neighborhoods-Atlanta-Analysis.ipynb

1. **Visualization using Folium**:

In order to have a first insight of Atlanta city venue types between neighborhoods, there is no better way than visualization.

The medium chosen is Choropleth map, which uses differences in shading or coloring to indicate a property's values or quantity within predefined areas. It is ideal for showing how differently real estate priced between neighborhoods across the Atlanta city map.

The map (Figure 3) shows various neighborhoods that located around Downtown, Midtown and greater Atlanta.

Downtown can be considered the heart of Atlanta city. It's where most businesses, tourist attractions and entertainments located. So, the venue types that can attract many people are expected to have the most positive coefficients in the regression model.

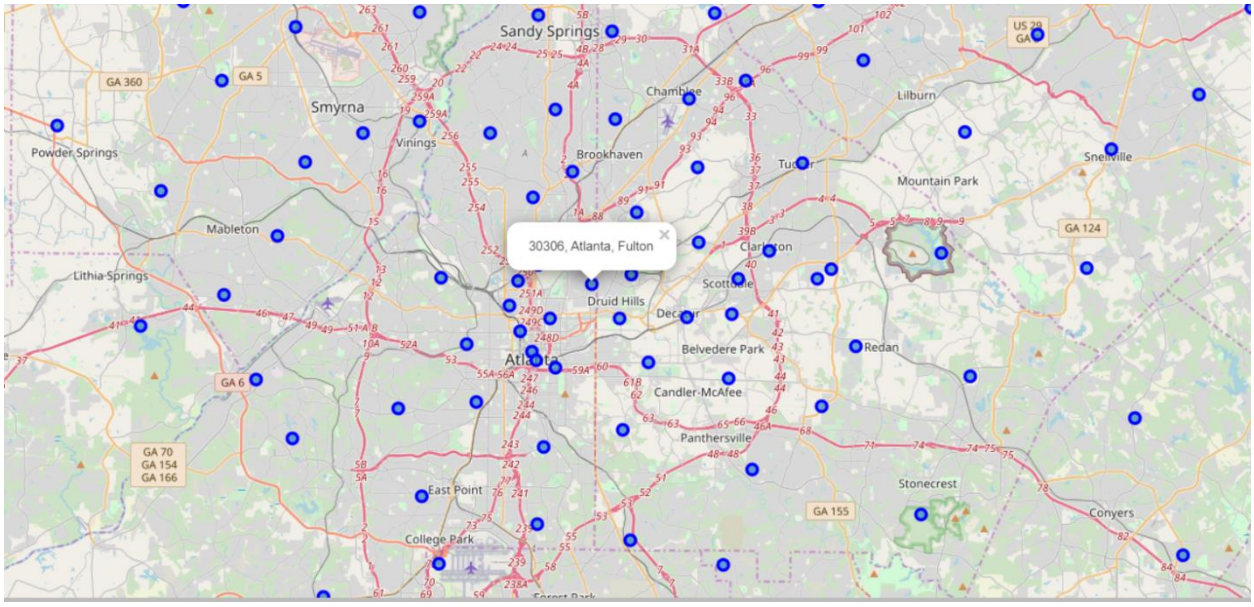

Figure 1 - Atlanta city neighborhoods in Folium Visualization

2. **KNN clustering and segmentation**:

Run *k*-means to cluster the neighborhood into 5 clusters

The result (Figure 3) shows the common venues in each neighborhood and it does seem very promising to start a new business in the area where the venue of interest may be missing.

[34]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Acworth	Farm	Farmers Market	Frozen Yogurt Shop	Fried Chicken Joint	French Restaurant	Football Stadium	Food Truck	Food Stand	Food Service	Food Court
1	Alpharetta	Home Service	Gymnastics Gym	Park	Construction & Landscaping	Food	Chinese Restaurant	Business Service	Martial Arts Dojo	Stables	Locksmith
2	Atlanta	Clothing Store	American Restaurant	Coffee Shop	Pizza Place	Mexican Restaurant	Park	Women's Store	Cosmetics Shop	Bakery	Department Store
3	Austell	Hookah Bar	Moving Target	Mexican Restaurant	Discount Store	Dive Bar	Frozen Yogurt Shop	Fried Chicken Joint	French Restaurant	Football Stadium	Food Truck
4	Avondale Estates	Home Service	Boutique	Thrift / Vintage Store	Gluten-free Restaurant	Yoga Studio	Food Court	Flea Market	Flower Shop	Food	Food Stand
5	Buford	BBQ Joint	Lingerie Store	Food	Furniture / Home Store	Frozen Yogurt Shop	Fried Chicken Joint	French Restaurant	Football Stadium	Food Truck	Food Stand
6	Clarkston	Grocery Store	Discount Store	Trail	Food Truck	Middle Eastern Restaurant	Football Stadium	French Restaurant	Food Stand	Farmers Market	Food Court
7	Conley	Office	Transportation Service	Yoga Studio	Fried Chicken Joint	French Restaurant	Football Stadium	Food Truck	Food Stand	Food Service	Food Court
8	Conyers	Gym / Fitness Center	Discount Store	Business Service	Construction & Landscaping	Yoga Studio	Food Court	Flea Market	Flower Shop	Food	Food Stand
9	Cumming	Sandwich Place	Harbor / Marina	Business Service	Construction & Landscaping	Miscellaneous Shop	Doctor's Office	Fast Food Restaurant	French Restaurant	Football Stadium	Diner

Figure 2 - Linear Regression result

Looking back further to the dataset, its dimensions sizes is clearly unbalanced, only 1000 venues returned by foursquare, and more than 300 features. Logical steps to take are either collecting more venues or trying to reduce the number of features.

But since there is no other public source available, increasing venue size is not possible now.

[55]:

	Latitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	33.797412	0	Bakery	Park	Automotive Shop	Convenience Store	Big Box Store	Campground	Lake	Garden Center	Pharmacy	Photography Lab
2	33.747999	0	Clothing Store	American Restaurant	Coffee Shop	Pizza Place	Mexican Restaurant	Park	Women's Store	Cosmetics Shop	Bakery	Department Store
4	33.794492	0	Clothing Store	American Restaurant	Coffee Shop	Pizza Place	Mexican Restaurant	Park	Women's Store	Cosmetics Shop	Bakery	Department Store
5	33.613639	0	Gym / Fitness Center	Discount Store	Business Service	Construction & Landscaping	Yoga Studio	Food Court	Flea Market	Flower Shop	Food	Food Stand
6	33.757576	0	Clothing Store	American Restaurant	Coffee Shop	Pizza Place	Mexican Restaurant	Park	Women's Store	Cosmetics Shop	Bakery	Department Store
7	33.791004	0	Clothing Store	American Restaurant	Coffee Shop	Pizza Place	Mexican Restaurant	Park	Women's Store	Cosmetics Shop	Bakery	Department Store
8	33.999280	0	Platform	Athletics & Sports	Garden	Moving Target	Basketball Stadium	Construction & Landscaping	Cosmetics Shop	Automotive Shop	Bank	Fried Chicken Joint
9	33.948631	0	Platform	Athletics & Sports	Garden	Moving Target	Basketball Stadium	Construction & Landscaping	Cosmetics Shop	Automotive Shop	Bank	Fried Chicken Joint
10	33.939949	0	Breakfast Spot	Cosmetics Shop	Hobby Shop	Speakeasy	Pool	Pharmacy	Football Stadium	Intersection	Arepa Restaurant	Bakery
12	34.124212	0	Dentist's Office	Southern / Soul Food Restaurant	Pizza Place	Chinese Restaurant	Pharmacy	Pub	Mexican Restaurant	Grocery Store	Shipping Store	Film Studio

Figure 3 – Atlanta Main cluster with top venues

Cluster 2 - Atlanta Sports Arena

[56]:
```
atlanta_merged.loc[atlanta_merged['Cluster Labels'] == 1, atlanta_merged.columns[[1] + list(range(5, atlanta_merged.shape[1]))]]
```

[56]:

	Latitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
57	33.97561	1	Food	Yoga Studio	Farmers Market	Frozen Yogurt Shop	Fried Chicken Joint	French Restaurant	Football Stadium	Food Truck	Food Stand	Food Service

Cluster 3 - Atlanta dining area

[57]:
```
atlanta_merged.loc[atlanta_merged['Cluster Labels'] == 2, atlanta_merged.columns[[1] + list(range(5, atlanta_merged.shape[1]))]]
```

[57]:

	Latitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
29	33.583798	2	American Restaurant	Yoga Studio	Farmers Market	Frozen Yogurt Shop	Fried Chicken Joint	French Restaurant	Football Stadium	Food Truck	Food Stand	Food Service

Figure 6 – Atlanta Sports arean and dining area cluster with top venues

Cluster 4 - Atlanta Farmers market area

```
atlanta_merged.loc[atlanta_merged['Cluster Labels'] == 3, atlanta_merged.columns[[1] + list(range(5, atlanta_merged.shape[1]))]]
```

	Latitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	34.034515	3	Farm	Farmers Market	Frozen Yogurt Shop	Fried Chicken Joint	French Restaurant	Football Stadium	Food Truck	Food Stand	Food Service	Food Court
93	34.103138	3	Farm	Farmers Market	Frozen Yogurt Shop	Fried Chicken Joint	French Restaurant	Football Stadium	Food Truck	Food Stand	Food Service	Food Court

Cluster 5 - Atlanta lake and recreational area

```
atlanta_merged.loc[atlanta_merged['Cluster Labels'] == 4, atlanta_merged.columns[[1] + list(range(5, atlanta_merged.shape[1]))]]
```

	Latitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
72	33.766909	4	Lake	Yoga Studio	Frozen Yogurt Shop	Fried Chicken Joint	French Restaurant	Football Stadium	Food Truck	Food Stand	Food Service	Food Court

Figure 7 – Atlanta Farmers market and lake area with top venues

These clusters provide a good insight into Atlanta neighborhoods and where a new business may be possible.

IV. Results:

Even though the neighborhood segmentation and clustering show high level information a more sophisticated analysis is needed, and the model is not suitable for specific prediction of business value. Thus, it can't be used to precisely predict a neighborhood venue of specific type and its value to the investor.

Explanations for the poor model can be:

- The neighborhood analysis in more detail is needed.
- The data is incomplete (small sample size, missing deciding factors).
- The machine learning techniques are chosen or applied poorly.

But again, on the bright side, the insight, gotten from observing the analysis results, seems consistent and logical. And the insight is business venues that can serve the needs of most normal people usually situated in pricy neighborhoods based on economic data of each zip code.

V. Discussion:

The real challenge is constructing the dataset:

- Many of the needed data isn't publicly available.
- When combining data from multiple sources, inconsistent can happen. And lots of efforts are required to check, research and change the data before merge.
- For data obtained through API calls, different results are returned with different set of parameters and different point of time. Multiple trial and error runs are required to get the optimal result.
- Even after the dataset has been constructed, lots of research and analysis are required to decide if the data should be kept as is or be transform by normalization or standardization.

It can be considered the most important process in the whole data science pipeline. Which can affect the most on the result.

On the other hand, choosing the suitable technique to construct the model is also a worthwhile process. As this report shows that, by applying a different method, the result can be improved.

VI. Conclusion:

In the short time allocated for this project, the analysis couldn't produce a precise model or showing a strong correlation for any venue type. But we can still get some meaningful and logical insights from the result.

Doing this project helps practicing every topic in the specialization, and thus, equipping learners with Data Science methodology and tools using Python libraries. Also doing a real project certainly helped me learn so much more outside the curriculum, as well as realizes what more to research into after completing the program. And as this report shows, there are surely a lot of things to dig into.

Many thanks for reviewing this project with your time and patience.

References:

k-nearest neighbors' algorithm

https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm