Venue Location Analysis using Foursquare and other Python’s libraries

Braulio J. Cabral

Coursera

06/14/2019

Abstract

[The abstract should be one paragraph of between 150 and 250 words. It is not indented. Section titles, such as the word Abstract above, are not considered headings so they don’t use bold heading format. Instead, use the Section Title style. This style automatically starts your section on a new page, so you don’t have to add page breaks. Note that all of the styles for this template are available on the Home tab of the ribbon, in the Styles gallery.]

Keywords: [Click here to add keywords.]

Venue Location Analysis using Foursquare and other Python’s libraries

# Introduction

A critical factor when opening a franchise is to determine if the location is suited for such venue. Other factors, such as distance from other franchises of the same corporation or competitor franchises are also important, as well as the need for people traffic through the location (Franchise Help, 2019). The business problem discussed in this paper relates to the need of a franchisee to validate a location for a new venue given by his corporate sponsor. Mr. John Doe received a proposed location in the town of Falling Waters, WV to open a fast food franchise. Event when the corporation’s suggestion looks appealing, he wanted to know more about the rationale, or come to his own conclusion to decide if he should proceed.

I proposed the use of Foursquare to collect a dataset of all businesses in the area and use data science techniques to analyze the dataset and see what we could learn from it. In the following sections of this paper, I will describe the dataset, the methodology and main components used in the exploratory analysis of the data and the resulting discovery.

## The Dataset

The proposed location is a small shopping center located at [39.56375, -77.887494]. The first approach in generating a data set was to map out all businesses in a 10-miles radius of this location. Businesses included all categories, competitors, such as fast food restaurants, and any other businesses that could serve similar menu items, such as small coffee shops or sandwich. places). To generate the dataset, at first, my approach was to use the above coordinates and run a single search. Unfortunately, that approach returned a number of businesses smaller than the actual number of business in the area. I then decided to list all coordinated of small towns in the area and loop through them to generate single outputs, then aggregate them into a super data frame. This approach resulted in a data frame with 368 business.

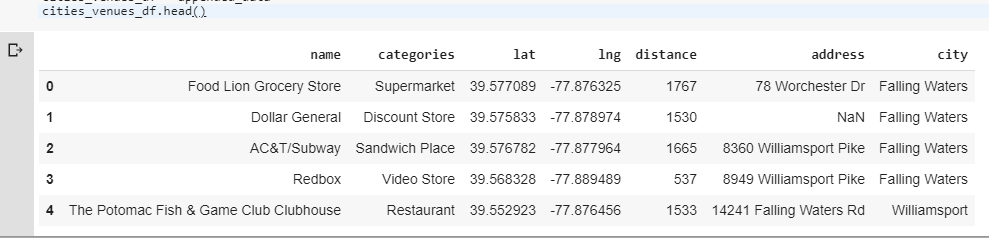


Fig. 1 (Data Frame of surrounding businesses)

The data frame above describe each venue as a row with the corresponding features of name of the venue, categories, latitude, longitude, the distance of the venue in relation to the proposed location, its address and the corresponding city or town. A very important feature here is the “distance” from the proposed location, for both, competing businesses, and non-competing. For competing businesses such as those of the same category as the new franchise, it will give us an idea of how strong the competition may or may not be in the proposed location, and for non-competing businesses, we can learn about the potential customers traffic in the area.

## Methodology

The first step in the exploratory data analysis of the above dataset was to produce a visual depiction of where the businesses are in relation to the proposed location. Using folium, a library the builds on the strengths of Python’s data wrangling capabilities and leaflet.js, I was able to produce a map and superimposed all businesses by their coordinates. To have a better view of where each corresponding town is, I overplayed each town as a colored circle enclosing the businesses.

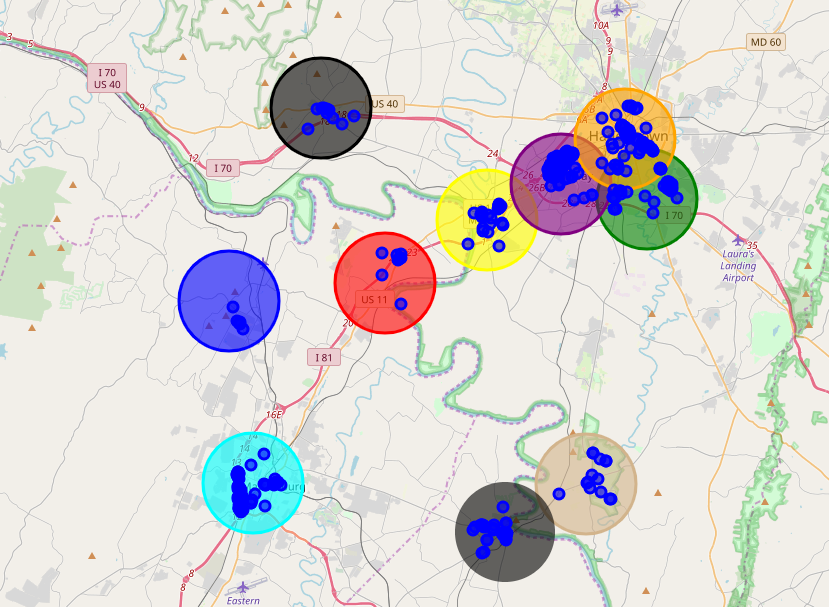


Fig. 2-Businesses locations and towns.

The map above shows how the businesses are naturally clustered within the different town’s boundaries. Later, we can see if using an algorithm such as KMeans produces a similar clustering. There were 112 unique business categories identified, the array below shows each category.



Fig. 3 -Array of businesses categories

As part of the exploratory data analysis, I use one-hot-encoding to group the different businesses by town, resulting in a data frame of 113 columns representing the business categories and 368 rows representing each business by town. A one-hot-encoding is a representation of categorical features as a binary vector (0 for non-present and 1 for present, in the case of the business categories). One-hot-encoding allows for a better representation of categorical data. For this problem, it helps visualize businesses considered competitors of the new franchise and those that can be potential generator of customers. For example, in the snip below we can appreciate that the town of Williamsport, MD does not have businesses that could be considered a competition to the proposed location given the fact that the franchise serves breakfast items. Of course, later in the exploration we can add the distance as a criterion to consider if this competitor is strong enough to have an impact on the new business (Brownlee, 2017).

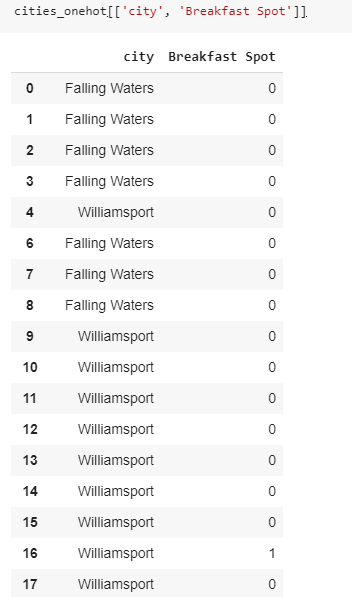


Fig. 4- Competitor identified using one-hot-encoding

Another technique used to analyze the categorical data was the distribution of 10 most common venues based in their frequency for each town. In the snip below, we can see that the closest direct competitors to the fast food franchise new location are in Williamsport, Hagerstown, Clear Spring, Hedgesville, and Martinsburg, and that Falling Waters itself does not have any fast food business, making it an attractive location so far.

----Falling Waters----

venue freq

0 Women's Store 0.14

1 Supermarket 0.14

2 Discount Store 0.14

3 Italian Restaurant 0.14

4 Sandwich Place 0.14

5 Video Store 0.14

6 Tourist Information Center 0.14

7 Playground 0.00

8 Pizza Place 0.00

9 Pharmacy 0.00

----Williamsport----

venue freq

0 Trail 0.11

1 Bar 0.07

2 Café 0.04

3 Breakfast Spot 0.04

4 Sandwich Place 0.04

5 Fast Food Restaurant 0.04

6 Pizza Place 0.04

7 Grocery Store 0.04

8 Discount Store 0.04

9 Pharmacy 0.04

----Hagerstown----

venue freq

0 Fast Food Restaurant 0.07

1 Pizza Place 0.06

2 Clothing Store 0.05

3 Sandwich Place 0.05

4 American Restaurant 0.05

5 Convenience Store 0.03

6 Pharmacy 0.03

7 Discount Store 0.03

8 Ice Cream Shop 0.03

9 Hotel 0.03

----Clear Spring----

venue freq

0 Restaurant 0.11

1 American Restaurant 0.11

2 Diner 0.11

3 Ice Cream Shop 0.11

4 Fast Food Restaurant 0.11

5 Hotel 0.11

6 Hot Spring 0.11

7 Food 0.11

8 Tea Room 0.11

9 Pet Store 0.00

----Hedgesville----

venue freq

0 Discount Store 0.17

1 Convenience Store 0.17

2 Pharmacy 0.17

3 Pizza Place 0.17

4 Fast Food Restaurant 0.17

5 Italian Restaurant 0.17

6 Video Store 0.00

7 Video Game Store 0.00

8 Pub 0.00

9 Playground 0.00

----Martinsburg----

venue freq

0 Pizza Place 0.10

1 American Restaurant 0.09

2 Mexican Restaurant 0.07

3 Fast Food Restaurant 0.05

4 Hotel 0.05

5 Coffee Shop 0.05

6 Italian Restaurant 0.03

7 Gas Station 0.03

8 Sandwich Place 0.03

9 Chinese Restaurant 0.03

Table 1 – Business frequencies by towns

We then reorganize the data by sorting the cities and the most common businesses, giving a quick way to see which are the most popular and for what categories. Again, here we can appreciate that Falling Waters most common venue is a women’s store, follow by Italian Restaurant, a donut shop that could be a direct competitor to the new business is in 8th place from the proposed location.

### KMeans Clustering.

KMeans clustering is one of the easiest to understand and simplest to implement unsupervised learning algorithms used in solving clustering problems (KM, 2019). For the purpose of this paper, I used KMeans to group businesses with similarities in different groups. Remember, the similarities will be based on the data passed to the algorithms, there is no ground truth provided, that is why this is considered an unsupervised learning algorithm.

I calculated the mean of the frequency identified in the one-hot data frame and passed the new data frame of business categories frequency means grouped by towns to the KMeans algorithm. But to determine the ideal number of clusters to pass in the K argument to the KMeans algorithm, I used agglomerative hierarchical clustering algorithm. This algorithm displays a tree-like diagram called a dendrogram. The end-tail of the dendrogram represents the possible nodes or clusters based on their dissimilarity or ward’s minimum variance in this case (NCSS, 2019).

The ward’s minimum variance is calculated as the coefficients of the distance of the equation shown below.

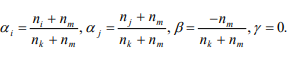


Fig. 5 – Ward’s minimum variance

The dendrogram suggests 10 nodes.

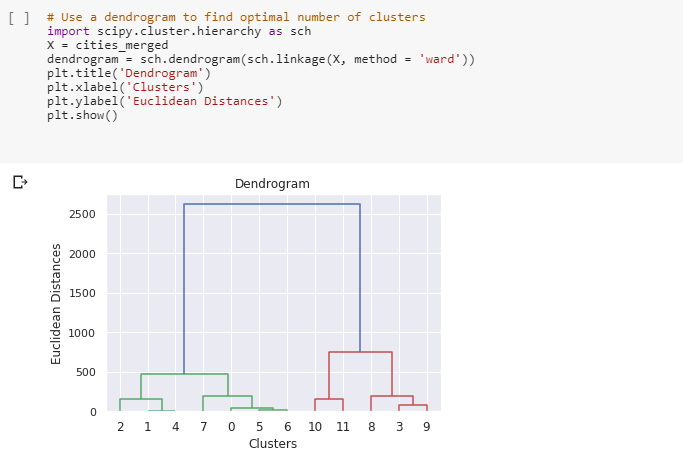
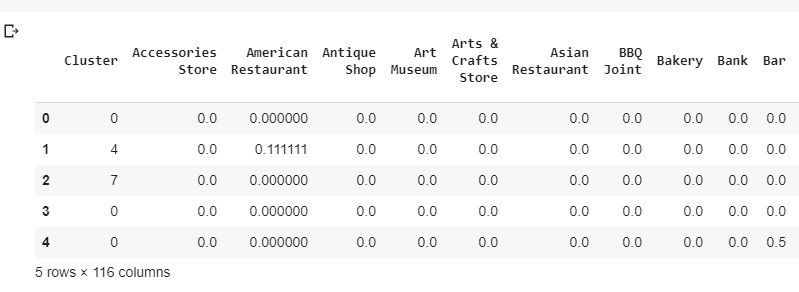


Fig. 6 – Dendrogram

Next, the data frame was further cleaned up to get rid of categorical data before feeding the frequency averages to the KMeans algorithm. Columns dropped included; name, categories, address, and city. Also, any rows with NaN in the latitude, longitude and distance was dropped, resulting in the data frame below.

….

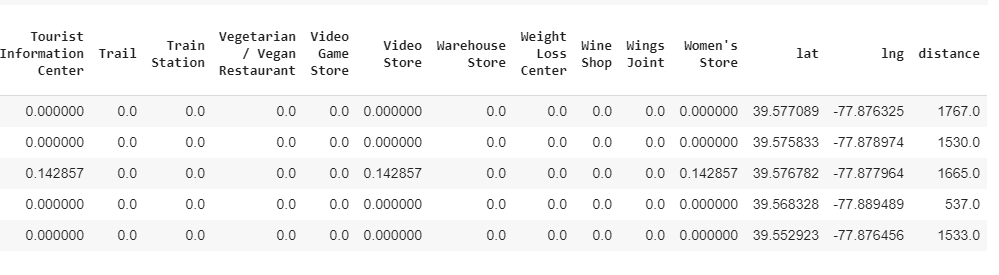


Fig. 7-Snip of data frame past to the KMeans algorithm

After transforming the data using StandardScaler() to ensure that values such as distance are standardized with the rest of the data. The standardized data is represented in the snip below in the form of an array.

array([[-1.14682928, 0. , -0.62286726, ..., -0.35199567,

-0.54912253, 0.87298433],

[ 0.10425721, 0. , 2.22125218, ..., -0.44304175,

-0.63828297, 0.45704071],

[ 1.04257207, 0. , -0.62286726, ..., -0.37423699,

-0.60428096, 0.69397062],

...,

[-0.52128604, 0. , 1.20549523, ..., 1.33941722,

1.29840416, -1.41909314],

[-0.20851441, 0. , -0.02758645, ..., 0.91614473,

1.80000389, -0.46084329],

[ 0.72980045, 0. , -0.62286726, ..., 1.36460766,

1.12535163, -0.74164911]]).

Table 2. Array of standardized data

[To add a table of contents (TOC), apply the appropriate heading style to just the heading text at the start of a paragraph and it will show up in your TOC. To do this, select the text for your heading. Then, on the Home tab, in the Styles gallery, click the style you need.]

### [Heading 3].

[Include a period at the end of a run-in heading. Note that you can include consecutive paragraphs with their own headings, where appropriate.]

#### [Heading 4].

[When using headings, don’t skip levels. If you need a heading 3, 4, or 5 with no text following it before the next heading, just add a period at the end of the heading and then start a new paragraph for the subheading and its text.] (Last Name, Year)

##### [Heading 5].

[Like all sections of your paper, references start on their own page. The references page that follows is created using the Citations & Bibliography feature, available on the References tab. This feature includes a style option that formats your references for APA 6th Edition. You can also use this feature to add in-text citations that are linked to your source, such as those shown at the end of this paragraph and the preceding paragraph. To customize a citation, right-click it and then click Edit Citation.] (Last Name, Year)

References

Brownlee, J. (2017). How to One Hot Encode Sequence Data in Python. Article extracted from URL: https://machinelearningmastery.com/how-to-one-hot-encode-sequence-data-in-python/

Franchise Help (2019). Where Is The Best Location To Open Your Franchise? Extracted from URL: https://www.franchisehelp.com/franchisee-resource-center/where-is-the-best-location-to-open-your-franchise-/

KM (2019). K-Means Clustering Algorithm. Article extracted from URL: https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm

NCSS Statistical Software (2019). Hierarchical Clustering/Dendrograms. Chapter 445. Extracted from URL: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Hierarchical\_Clustering-Dendrograms.pdf

Last Name, F. M. (Year). Article Title. *Journal Title*, Pages From - To.

Last Name, F. M. (Year). *Book Title.* City Name: Publisher Name.

Footnotes

1[Add footnotes, if any, on their own page following references. For APA formatting requirements, it’s easy to just type your own footnote references and notes. To format a footnote reference, select the number and then, on the Home tab, in the Styles gallery, click Footnote Reference. The body of a footnote, such as this example, uses the Normal text style. (Note: If you delete this sample footnote, don’t forget to delete its in-text reference as well. That’s at the end of the sample Heading 2 paragraph on the first page of body content in this template.)]

Tables

Table 1

[Table Title]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Column Head | Column Head | Column Head | Column Head | Column Head |
| Row Head | 123 | 123 | 123 | 123 |
| Row Head | 456 | 456 | 456 | 456 |
| Row Head | 789 | 789 | 789 | 789 |
| Row Head | 123 | 123 | 123 | 123 |
| Row Head | 456 | 456 | 456 | 456 |
| Row Head | 789 | 789 | 789 | 789 |

Note: [Place all tables for your paper in a tables section, following references (and, if applicable, footnotes). Start a new page for each table, include a table number and table title for each, as shown on this page. All explanatory text appears in a table note that follows the table, such as this one. Use the Table/Figure style, available on the Home tab, in the Styles gallery, to get the spacing between table and note. Tables in APA format can use single or 1.5 line spacing. Include a heading for every row and column, even if the content seems obvious. A default table style has been setup for this template that fits APA guidelines. To insert a table, on the Insert tab, click Table.]

Figures title:

Figure 1. [Include all figures in their own section, following references (and footnotes and tables, if applicable). Include a numbered caption for each figure. Use the Table/Figure style for easy spacing between figure and caption.]

For more information about all elements of APA formatting, please consult the APA Style Manual, 6th Edition.