Exploring Ivy League universities neighboring locations using Foursquare API

(IBM Data Science Professional Certificate Capstone)

Sergiy Tkachuk

January 13, 2019

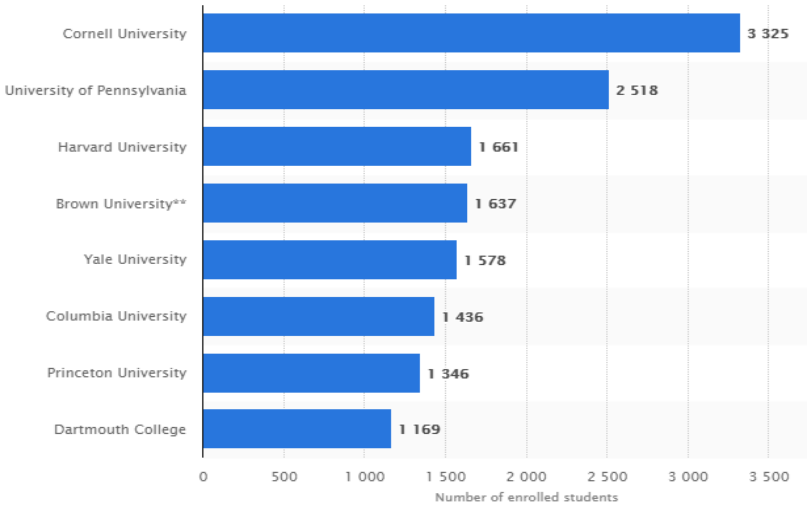
1 Abstract

Right choice of the location is one of the most import steps while starting up a ne a new business. The location can be determined by target audience, potential number of customers and therefor greater revenue and turnover. Each place has its own characteristics that are significant for deciding about business localization. This project concentrates on exploratory analysis on set of different venues collected from different data sources. In conjunction with machine learning clustering methods it attempts to identify right place to establish new business by its type.

2 Introduction

An artificial business case for the study could be formulated as follows: *how to help US based grocery and restaurant chain to expand in Northeast United States?*

Presence in this part of the US will significantly increase chains brand recognition and positively impact overall sales in the country. What is more it will allow fight competition in target states. This motivates the company to use advanced analytics that can help to take right decisions. First step during the expansion is to conduct a research on local market and potential competition.



Company has decided to target student first, because of Ivy League universities presence in the area. That could potentially boost the acknowledgement of the brand and bring more customers as student represent a huge portion of the region population.

This report is illustrating the analysis of neighboring coffee shops, bars, shops and pubs. The exploration is run mainly to evaluate potential of new business around the universities. First half of the report is dedicated to data extraction and building working data set, whereas the second part is about project results, conclusion and open point for discussion.

3 Data

As the first step the list of Ivy League universities is gathered from Wikipedia page. In this study Foursquare API is used as a primary source of information to extract nearby venues data. In particular longitude and latitude are used to search for different locations. Python **geopy** library was used to define geographical coordinates for universities by their names.

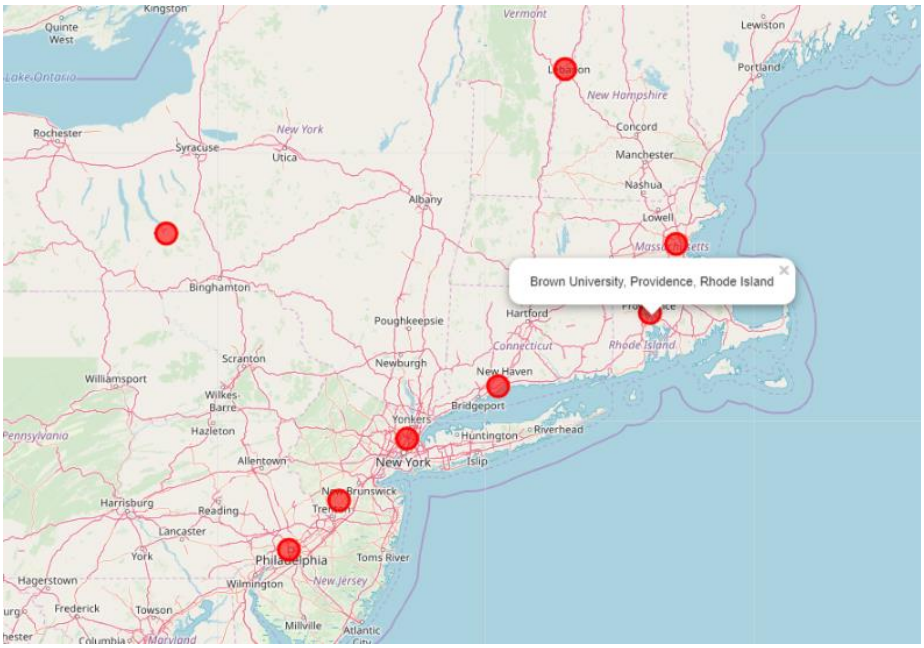
Institution	Location	Athletic nickname	Enrollment	2017 endowment	Academic staff
Brown University	Providence, Rhode Island	Bears	8649	$3.5 billion	736
Columbia University	New York City, New York	Lions	22920	$10.0 billion	3763
Cornell University	Ithaca, New York	Big Red	20633	$6.8 billion	2908
Dartmouth College	Hanover, New Hampshire	Big Green	6141	$4.96 billion	571
Harvard University	Cambridge, Massachusetts	Crimson	21225	$37.1 billion	4671
University of Pennsylvania	Philadelphia, Pennsylvania	Quakers	20643	$12.2 billion	4464
Princeton University	Princeton, New Jersey	Tigers	7592	$23.8 billion	1172
Yale University	New Haven, Connecticut	Bulldogs	11666	$27.8 billion	4140

4 Methodology

Some research say that it is good for a competitive business to be located near other businesses of the same type.

I used Foursquare API search and explore features to investigate venues close to each university. The output of the API call is being returned in json format, so I converted it to pandas DataFrame to be able to run further analysis. Feature engineering and other exploratory analysis were done afterwards.

Spread of the found venues are pictured using **folium** library. Map with highlighted universities is shown below:



For the sake of simplicity, the following analysis is performed for two universities: Cloumbia University and Brown University.

By exploring each venue, the working data set has been enriched. The data that has been received from Foursquare API looks as follows:

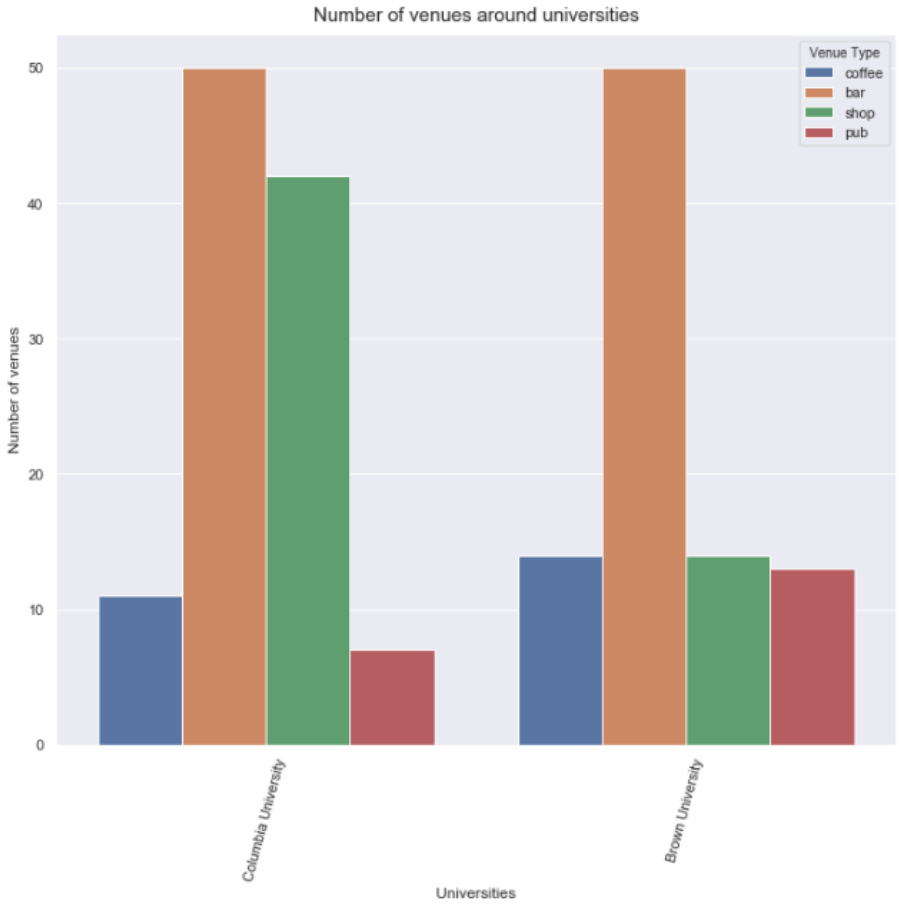
id	name	city	state	categories	verified	likes	rating	photos	tips	open_days	avg_hours_open	comments
4b057f84f964a520d15822e3	Blue State Coffee	Providence	RI	Coffee Shop	0	112	8.0	43	27	7	10.0	Take a token from the counter and drop in the ...
4b0586aaf964a520816922e3	The Coffee Exchange	Providence	RI	Café, Coffee Shop	0	162	8.5	83	70	6	11.5	This is what the energy of a coffee shop shoul...
53bc1aee498e6b2c8c152021	Dave's Coffee	Providence	RI	Coffee Shop	0	76	9.1	37	18	7	10.1	Great atmosphere, great interior design and fe...
4ad78fc2f964a5202d0c21e3	Starbucks	Providence	RI	Coffee Shop	0	73	7.2	33	16	6	10.0	IF THE DRINK I CALLED IS NOT WHAT YOU ORDERED,...
4a5e0d87f964a520f6bd1fe3	Starbucks	Providence	RI	Coffee Shop	0	63	7.5	64	21	6	11.0	try the wifi, get it unsweetened.

Comments has been transformed to vector representation by using TfidfVectorizer.

5 Results

The chart below id showing the distribution of different venues type located around selected universities. As we can see bars are clearly dominating in both schools, while there is big discrepancy between number of shops and bars. Coffee shops are nearly the same level.

Intuition tells me that it is not probably a good idea to open bar neither near Columbia University, nor Brown University. What could be interesting is opening pub near the former and shop near the later. Let's validate the assumption by conducting the study.



The results of clustering are showing that it is better to open café in both location. Assuming that we use only two clusters (e.g. grocery and restaurant) it makes sense. Exploration of the produced clusters proves this.

6 Discussion

There are some open point worth discussing and maybe further study. One of them is trying to identify clients/customers by persona. Which means to run the analysis on the most granular level and gathering user data. That is labor-intensive and time-consuming task. Legal concerns should be taken into account as well here.

Other point is to perform regression analysis of potential sales and price positioning in selected areas. That might have significant impact on decision making process as in directly influence basic aim of the company – maximize its revenue.

Last but not least id to go through the offer of the competition and compare it to the market trends. Having this done the company could be ready to launch some brand-new offering keeping in mind that others does not have this, and it will take considerable amount of time to adjust their offer. In the meantime, the company can take over big part of the local market.

7 Conclusion

Driving the conclusion from the study we can say that it is worth opening café near campuses of both Columbia and Brown Universities. The reason behind the conclusion is of course limited to the data used. All thoughts and concerns are explained in previous section.

References

[1] Foursquare Developers Documentation, https://developer.foursquare.com/docs

[2] Ivy League members data, https://en.wikipedia.org/wiki/Ivy_League

[3] IBM Data Science Professional Certificate course materials, https://www.coursera.org/specializations/ibm-data-science-professional-certificate

[4] Ivy League Stats, https://www.statista.com/statistics/941545/ivy-league-students-class/