IBM – Coursera

Data Science Specialization

Capstone project - Final report

**The Best location to open a coffee shop in Boston Area**

Mehmet I Erdogdu – 2019

# Introduction

Boston is the capital and most populous city of the Commonwealth of Massachusetts in the United States. The Boston area has many colleges and universities make it an international center of education including law, medicine, engineering, and business, and the city is considered to be a world leader in innovation and entrepreneurship, with nearly 2,000 startups. Boston's economic base also includes finance, professional and business services, biotechnology, information technology, and government activities. Coffee shops are a popular place for informal business meetings or for students to catch up on schoolwork. Walk into any popular coffee shop and chances are you will see a group of students collaborating on a school project.

The mail goal will be focusing on analyzing universities and coffee shops data in Boston area to be able to find the least busy coffee shop in terms of student numbers per Coffee Shop.

Having a great location is obviously very important for attracting customers when you're opening a coffee shop. For that reason I will be focusing on the areas that is closer to the Universities and also I will be looking the number of coffee shop location for each university for a random radius. I will also be considering the student population for each university.

To make the choice of best location, the following information is needed:

-List of the universities in Boston area

-List of the coffee shop in Boston area

-The number of the student for each university

# Data

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

Wikipedia, which provide the Enrollments of each university in State of Massachusetts. <https://en.wikipedia.org/wiki/List_of_colleges_and_universities_in_Massachusetts>

FourSquare API which provides the surrounding venues of a given coordinates.

The process of collecting and cleaning data:

* Scrap the Wikipedia webpage for a list of Massachusetts universities and enrollment numbers for each university

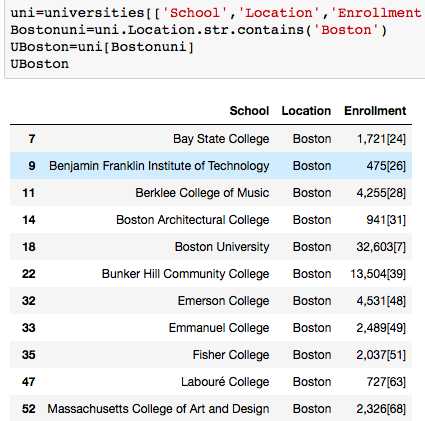


Figure 0.1 List of Universities and Enrollment numbers

* Then, Foursquare tool will be necessary to explore the coffee shops around the each University. By using API queries, it is possible to aggregate coffee shop location data.

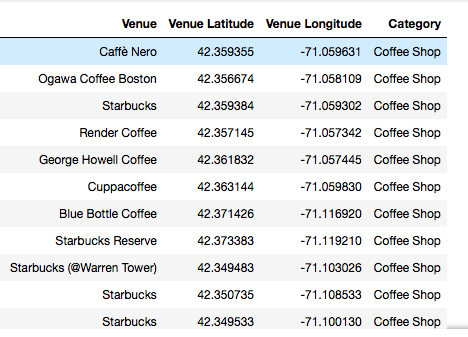


Figure 0.2 List of Coffee Shop location

# Methodology

By direct appreciation of the neighborhoods of the Universities in Boston, it is decided to perform the analysis on the district of Boston as an objective.

After retrieving the coffee shop location data and combining them with University data, it will be possible to manipulate the data accordingly

Once the objective of the study has been determined, marking and grouping techniques are put into practice, which will be necessary to analyze the best options and make a decision based on the data obtained. The procedure followed will be the same as that already concluded in the case study of the city of New York. Points of interest will be searched in the designated neighborhood and a grouping will be made by points of interest that will help us decide which steps to follow.

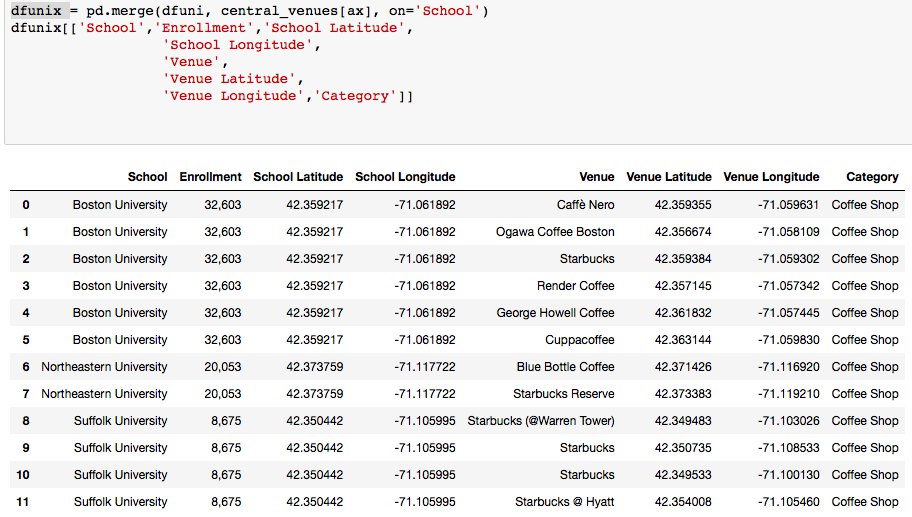


Figure 0.3 Coffee Shop locations near Universities

# Results

After manipulating all the data, I finally discovered that there are 3 University in Boston area. I calculated total number of the coffee shops with the radius of 500. As you can see in the table below there are 6 coffee shops near Boston University, which has 32,603 students currently enrolled. That means 5,434 students per coffee shop. If I continue calculating the other two shops, near Northeastern University 10,027 student and near Suffolk University 2,169 student per coffee shop.

As we can see, it is better to open a coffee shop close to the Northeastern University since it has the most student per coffee shop with average 10,027 students.

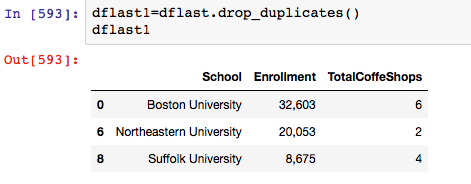


Figure 0.4 The Universities and total Coffee Shops within 500 ml

# Discussion and Conclusion

The real challenge is constructing the dataset:

* Usually 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.

The case study will be very useful for entrepreneurs and could be expended to other venue categories, as other correlations could be identified and explored using statistical inferences and Data Science techniques.