Finding the Best Location to Start a New Business

Denis Silva

November 20, 2019

1 Introduction

1.1 Background

Starting a new business is challenging and requires time and effort, more importantly, entrepreneurs need to determine how they're going to cover costs, and if they have the means to fund them. In order to make profit and stay afloat, they always need to be growing their business and a good spot can contribute to the success of the business.

When looking for space to house the business we need to make sure: what type of location is best; how much is the rent; and, if the proposed spot is appropriate for the business. Location for companies that have great contact with the public on their premises is critical, it must be defined how customers will visit the business, it will be on foot, public transportation (bus, train or subway) or private vehicle, if so, is there enough parking space? and will the parking be for free? These are all questions that should be carefully planned.

A crowded location where renting is high does not necessarily mean that the customer volume will increase. In short, business space must be carefully defined in order to reach the target customers.

Finally, renting has a tremendous impact on the business' finances, especially during the first year, so picking a spot becomes decisive. Entrepreneurs should spend as little as possible when starting and only on things that are essential for the business to grow and be successful. An *Electronic Repair Shop* fits this profile, so it's necessary to study all scenarios to better predict a location.

1.2 Problem

Identify the best location to start a new business in Boston which is far from downtown. The target customer is people living nearby the store and that suffer traveling a long distance to the city center to have just fixed their devices. The easy access by public transportation is a big concern as the in-home service can take advantage of it.

Find a location that can boost the company's services is a desire, in addition, partnership can help bond with customers and improve profits. Collaborating with more established brands in your industry is a great way to achieve growth. Reach out to other companies or even influential bloggers and ask for some promotion in exchange for a free product sample or service. Partner with a charity organization and volunteer some of your time or products to get your name out there.

1.3 Interest

Entrepreneurs investing in new businesses that want to find the best location that helps booster their business. Also, Advisory Companies that support entrepreneurs in their decisions to start business and plan their future.

2 Data

This section discusses data acquisition and wrangling processes - responsible for data cleaning and preparation. These phases are crucial to get accurate results for our customers.

2.1 Acquisition

Most of the data come from the official Boston website (https://data.boston.gov) where we will extract census and geolocation. Subway station locations were downloaded from https://opendata.arcgis.com. The geolocation demarcates neighborhood limits onto a map. Census information accounts for the population in each neighborhood along with its the age range.

All data made available on Boston website was in the right format – csv, xlsx and geojson – to be explored, resulting in an easy merge of data.

Foursquare API returned all companies around each candidate area along with their statistics such as the number of likes, ratings, and so on. We expect that statistics information help us get better predictions and figure out citizens profile. Next, will be categorize companies accordingly their type – competitors: direct competitors, general repair shop: company which does general repair and is not a rival, finally, common: companies which do not fit the preceding categories. Candidate areas were produced using a grid system over Boston's map with downtown excluded (radius of ~5Km from the city center).

2.2 Wrangling

After a radius of approximated 5Km from the city center had been excluded, only 22 neighborhoods remained to be studied. Census information for each neighborhood was merged with two age range in order to understanding pattern of young and old population. Regarding address and geolocation, less than 5% of addresses around candidate areas hadn't its address identified. Finally, subway stations' geolocation were used to identify proximity of predicted locations. Public transport is well-distributed in Boston, what made unnecessary calculate their distance from candidate areas.

Foursquare API did not give access to all statistics data about registered companies such as number of checkins, visits and chains. Though, a great number of missing values was reported and replace by zero. Those data were extremely valuable to Exploratory Data Analysis identify patterns and strong correlations. Also, those were the only missing values found.

We have only gathered detailed information for competitors and booster companies due to free account limitations.

The major difficulty was missing statistics data from companies, the fallout of this lack has reduced statistics conclusion.

2.3 Features selection

The main data set was composed of 1635 samples and 22 features, "store_type" was created later to differentiate the four types of stores – competitor, general, booster, and common. The categories for each store type is as follows: 7 booster neighbors, 2 competitors, 3 general repair, 236 common, totaling xx unique categories. In fact, competitors represent only 0.98% percent from total stores obtained from Foursquare which was: 16 competitors, 13 general repair and 1609 common.

Company statistics – likes, ratings, number of photos, is it a chain, price, url - was obtained only for competitors, general repair and booster neighborhood, due to plan limitation, totaling 35 samples. Company statistic is composed of 9 features. The correlation study shown that only 4 features meaningful value for the study, all remaining were removed. The features "id" and "url" were unnecessary to the analysis, likewise "price" which is null for all registers. Although, "cartesion_x"

and “cartesian_y”, known as UTM, are the same information that “latitude” and “longitude”, known as WGS86; UTM coordinates are better for calculating the distances.

A correlation study was applied and identified that the cartesian values and latitude & longitude produced a strong correlation (>.90) as expected. Other meaningful correlations found were: number of photos and ratings, meaning that as the company improves the quality of its services the rating also increase; tips and likes, meaning the growth in the number of tips results in more likes; the number of photos has relations with likes such as a before-and-after image could gain many likes. The feature ‘Chain’ could expose some useful insights, but, it was completely null.

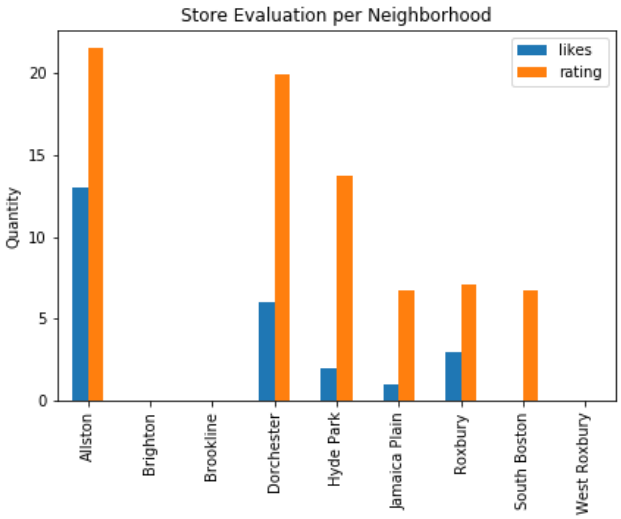
The data samples finished the cleaning phase 10 features.

3 Exploratory Data Analysis

Here, we evaluate correlation and pattern in order to get insights about customers and stores, especially, understand how our competitors are distributed over Boston.

Relationship between age and online evaluation

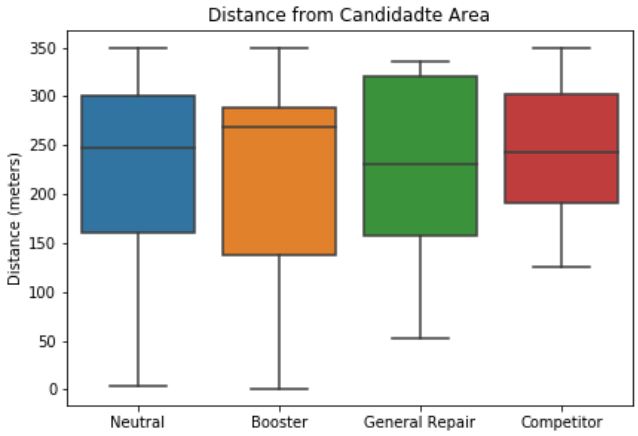
It is widely known that young people are highly active with online community, they share much more information than older ones. Allston has the youngest population in Boston, with 67% of them in the range of 20 to 34 years. The second neighborhood is Brighton with 56% of the population. Allston-Brighton is considered as one region due to its size and approximation. In the following figure, Brighton numbers are null because it does not have any General or Electronic Repair shops.



Relationship of distance from candidate areas and company type

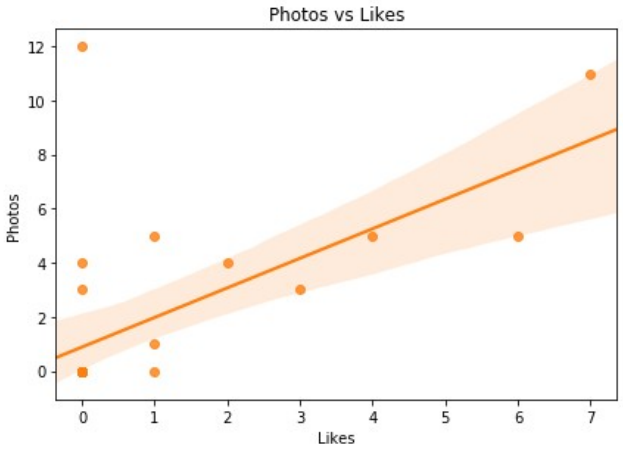
We limited our radius to 350 meters for each candidate area. As the following image shows, this radius limit was respected. The median for all store types have approximately the same value, with 250 meters of distance from candidate areas. Competitors are more distance, its closest store is 150 meters away from the closest candidate area.

We can conclude that all store types, except for competitors, have 25% of its companies up to 150 meters to the candidate areas.



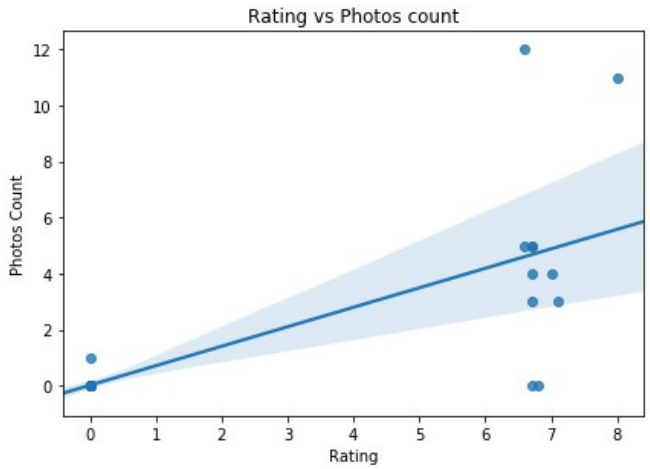
Relationship between the number of likes and photos online

The correlation between likes and photos is 0.62, it may be stronger if more data was available. The number of likes can increase if there are more posted photos. Photos such as before-and-after of services, store environment and nearby area can attract more customers. There are a few outliers with a greater number of images and zero likes, the lack of “statistic” data about companies could help to infer the discrepancy.



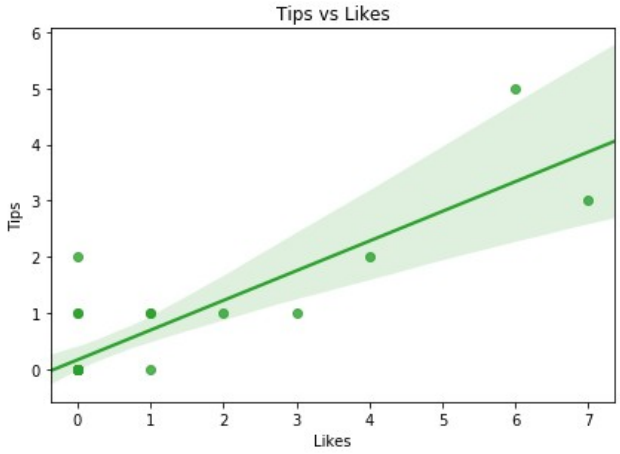
Relationship between rating and photos quantity

A correlation of 0.73 for photos and rating is possible to infer that as the company increases its quality the number of photos also increases, these photos could be before-and-after images from customers. A further step is to deep dive in images analysis and understand its type.



Relationship between the number of tips and likes

It is clear this pattern where the number of likes increases when more tips are given. Customers are more inclined to accept tips, suggestions, and orientations from other ones. It is human nature trust in other customers' opinions, so giving orientations and explaining about problems could help boost customers' satisfaction and open new knowledge. An updated blog with tips, orientation, and news cold step up popularity.


Tips vs Likes
Tips
Likes
0
1
2
3
4
5
6
7

4 Predictive Modeling

The clustering model will be used to predict the best location to start a new business. It computes the optimal centroids (central point inside an area) based on the number of clusters we are working with. An elbow curve is a chart that supports our decision about the ideal number of clusters for the model. As shown in *figure 4.1*, after 8 clusters the prediction stops to improve, so, we will study 6 and 8 clusters predictions.

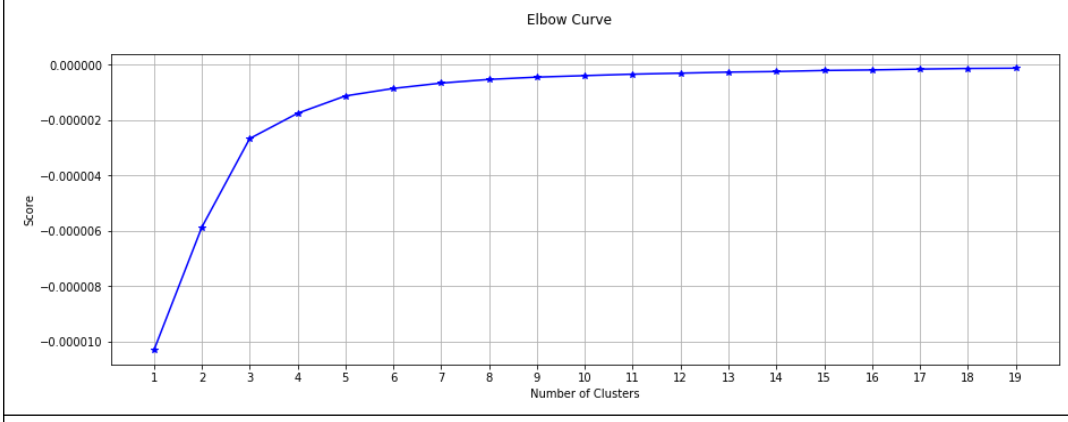


Figure 4.1 – Elbow Curve

We extracted latitude and longitude from competitors, general repair shops, and booster neighbors and used as input for the K-means model. Then, we compared its resulting regions with the locations that had been manually defined using business constraints and requirements. After we had mapped all centroids, we plotted onto a map with a radius of 300 meters around each one. The following sections detail the results we have achieved. And, to better understand these achievements we must recognize the following points:

- **Green circles:** k-means predictions, that is, clusters predicted;
- **Purple circles:** subway stations;
- **Blue circles:** location close to competitors; (aimed location)
- **Black circles:** location close to booster neighbors.
- **Heat map mark:** represents competitors (southeast Dorchester) and booster neighbors stores (remaining marks).

The terminology manually defined location is frequently used here, it represents all locations that we had defined based on business restrictions and requirements.

4.1 K-means using 8 clusters

In this test we have defined 8 clusters (k=8), that is, we expect the model produces eight regions. This simulation has produced a good result where two of our manually defined locations were inside the centroid radius. The “aimed location“ (blue circle) was partially inside a prediction, while the booster neighbors (black circle) had the best prediction, completely, inside centroids.

An extract region in which we manually defined 4 possible addresses, south Dorscherster, Dorchester-Milton limits, and 2 Boston’s neighborhoods. The figure 4.2 shows more details about this achievement.

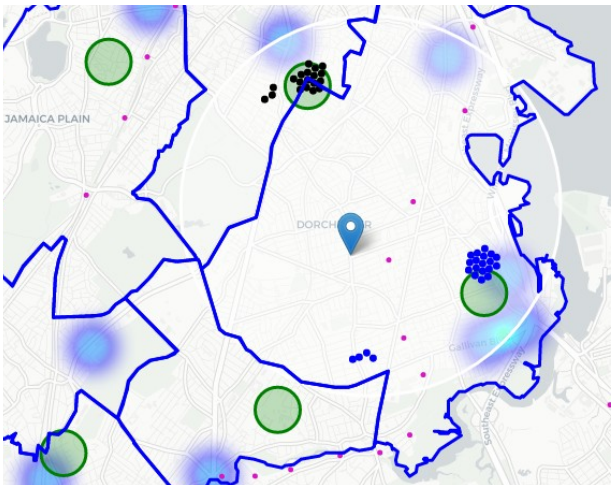


Figure 4.2 – 8 Cluster

4.2 K-means using 6 clusters

In this test were used 6 clusters (k=6), that is, we expect the model generates six ideal locations. The “aimed location” (blue circle) produced a perfect result, it was completely inside the centroids. By this time, booster neighbors are far from centroids.

Lastly, four addresses located at south Dorchester which are near to subways a station are now closer to a predicted centroid.

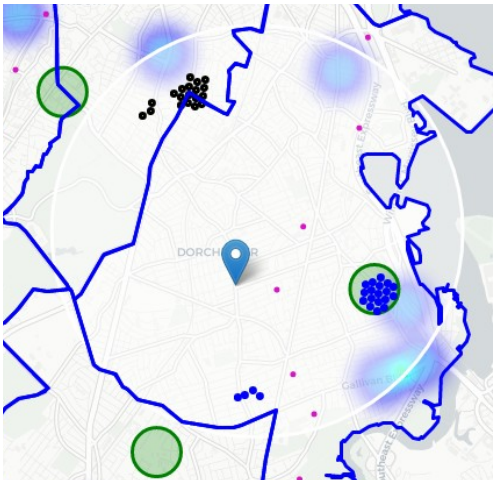


Figure 4.3 – 6 Cluster

4.3 Map Location

This last section shows the predicted locations over the Boston map. We focused Dorchester region. The *figure 4.4* is the first image shown, it is zoomed in to get more details of the “aimed location” (blue circle). The mark in the middle is Dorchester center.

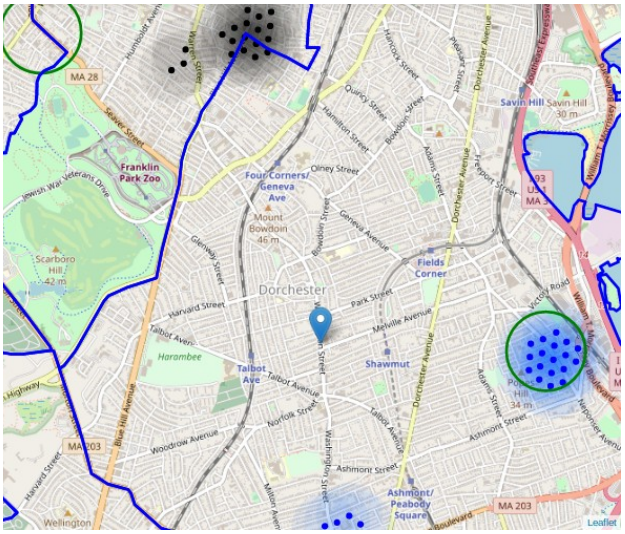


Figure 4.4 – Dorchester Zoomed

The *figure 4.5* shows the predicted location at Dorchester neighborhood zommed out, in order to get a full relation of other centroids.

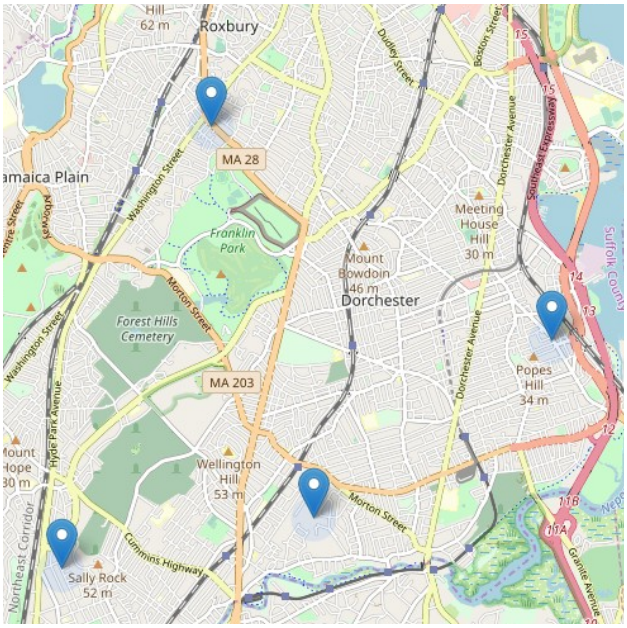


Figure 4.5 – Predicted Locations

5 Conclusion

This study analyzed the Boston area and companies in each neighborhood to try to predict the best spot to start an electronics repair shop. Besides, during the search for the best location, we discovered that photos, likes, and ratings were correlated with the number of likes and tips. A second breakthrough was that public transportation won't be a problem as it's well distributed and available throughout Boston.

Our manual filtering and clustering model have predicted Dorchester as the greatest candidate region to start a new business. Although Dorchester high probability, Allston had good results too. But, the lack of information about its local business and surrounding cities – that weren't included here – reduced its chances. Allston and Brighton be considered as one region, fact that increases, even more, its chances.

Finally, the entrepreneurs should define a marketing plan to launch and build a clientele, reaching the greatest number as possible by getting the word out about their business. This process, in the beginning, is just as important as choosing the best location. The correlation makes this conclusion clear.

6 Future directions

More data about companies' statistics such as likes, ratings, and so on, must be gathered. A Foursquare premium plan can unlock them. As reported, those statistics are fundamental to correlation and empower our predictions. Although the small amount of data – about competitors and similar business - we were able to identify patterns like Allston neighborhood be a good candidate region, even its small population, its surrounding cities could be supporting its commercial area. As result, we should expand this study to surrounding cities – Cambridge, Chelsea, Newton, Brookline, Watertown, Quincy, and Milton – in order to get meaningful insights regarding their impact on Boston local business.

Also, we can try to identify the appropriate type of spot, such as the type of building facilities and if the business space is adequate. And, study Boston renting neighborhood will, definitely, help to narrow down locations.