Effects of Public Transportation on Businesses

Michael Pretko

March 7, 2020

Introduction / Business Problem

In many major cities, such as New York and Boston, a large fraction of the population gets around primarily via public transportation, particularly the subway system. A natural question is whether proximity to public transportation plays an important role in deciding the success or failure of a business. Should distance to subway stations be a key factor in deciding the location of a new business? This information is crucial for people seeking to start new businesses in dense urban areas.

To this end, I will analyze how proximity to public transportation affects the success of businesses in a major city. I focus on Boston, where I lived for many years. I will see how various metrics of success, such as the ratings and number of likes of business venues, depend on their distance from the nearest subway station. Does proximity to public transportation help or hurt businesses? Does the answer depend on the particular type of business, or is it fairly universal? And is this correlation strong enough to make access to public transportation an important factor in deciding the location of a new business? By analyzing location data and ratings metrics obtained from Foursquare, I will produce actionable intelligence which will be highly valuable to prospective new businesses.

Data

For this project, I require data about various business venues, as well as public transportation data. I will also focus on metro (subway) data, since this is the primary means of public transportation in Boston. Using the category ID for metro stations on Foursquare, I started by identifying all metro stations in the urban core of Boston (defined as within three kilometers of the city center). I limited the search to the 30 most used metro stations, which helped to eliminate redundant information (e.g. duplicate Foursquare entries for the same station). The locations of these metro stations are visualized in Figure 1.

I then obtained data from a large sample of business venues throughout the city, by searching for the top 20 venues within one kilometer of each metro station. (This produced a number of duplicate entries, which needed to be eliminated at the data-cleaning stage.) I obtained the location, ratings, and number of likes of each of these venues. Note that ratings and likes are obtained via premium Foursquare calls, which limited the overall size of the data to 500 venues.

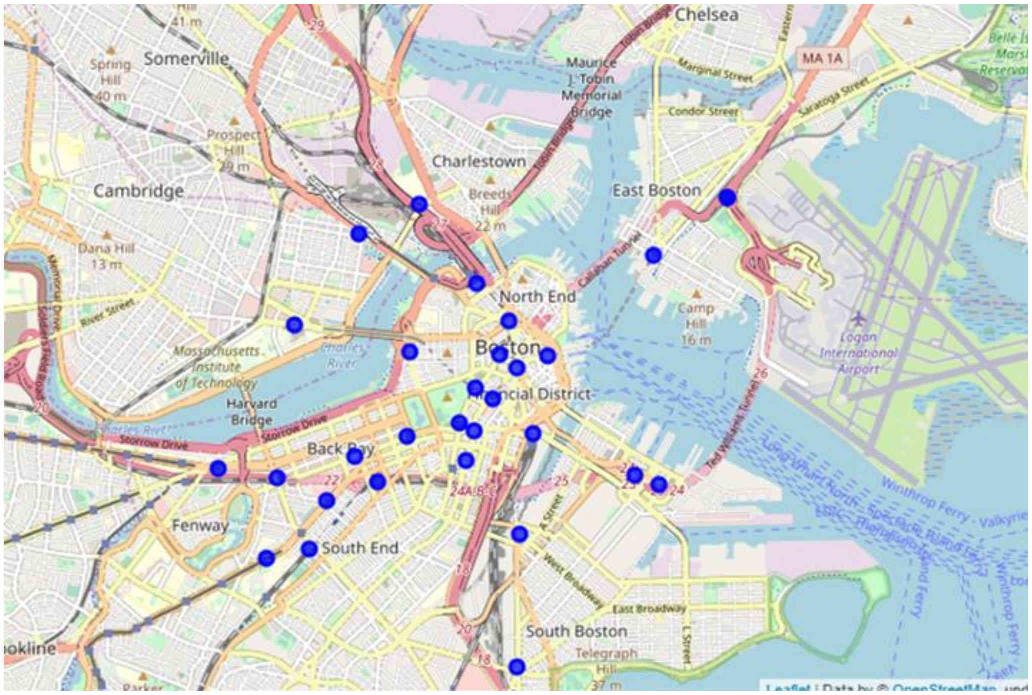

Figure 1: Map of metro stations in Boston

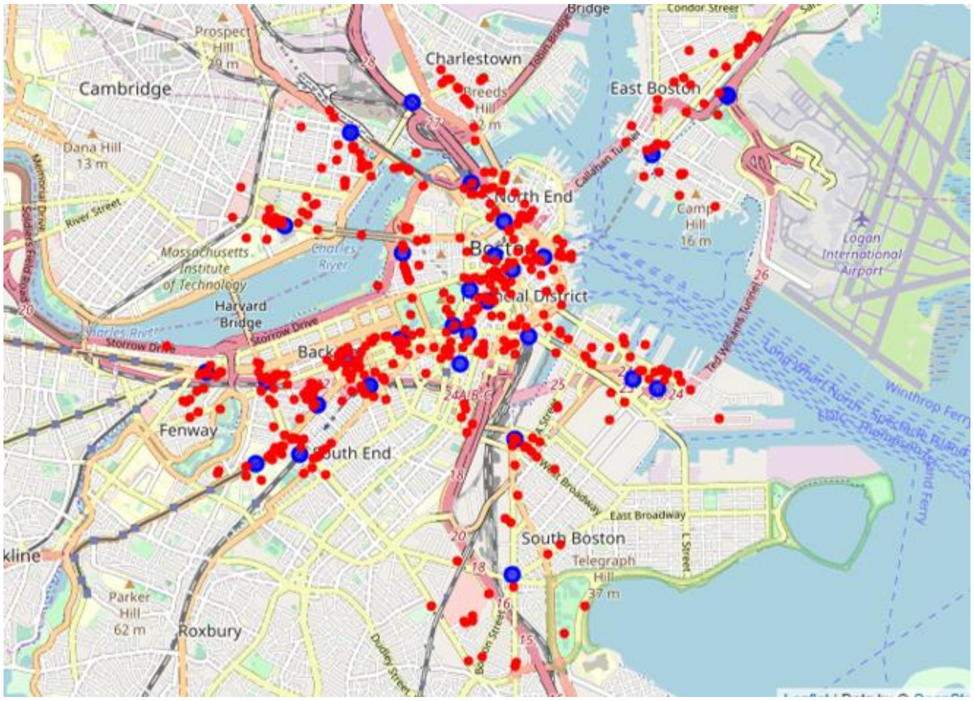

Figure 2: Map of venues and metro stations in Boston

Using the locations of both the metro stations and the venues, I used the geopy library to calculate the distance of each venue to the closest metro station. The locations of all venues are shown in Figure 2. As can already be roughly seen from the map, the top venues seem to be clustered fairly close to the metro stations (with a few notable exceptions). To get more concrete results, I will analyze all datasets to see what (if any) correlation there is between success of a business and proximity to public transportation. I will then use this information to produce recommendations for people starting new businesses.

Methodology

As a first step in understanding how distance from public transportation affects businesses, I created a histogram showing how many venues are in certain ranges of distances away from their nearest metro station, as depicted in Figure 3:

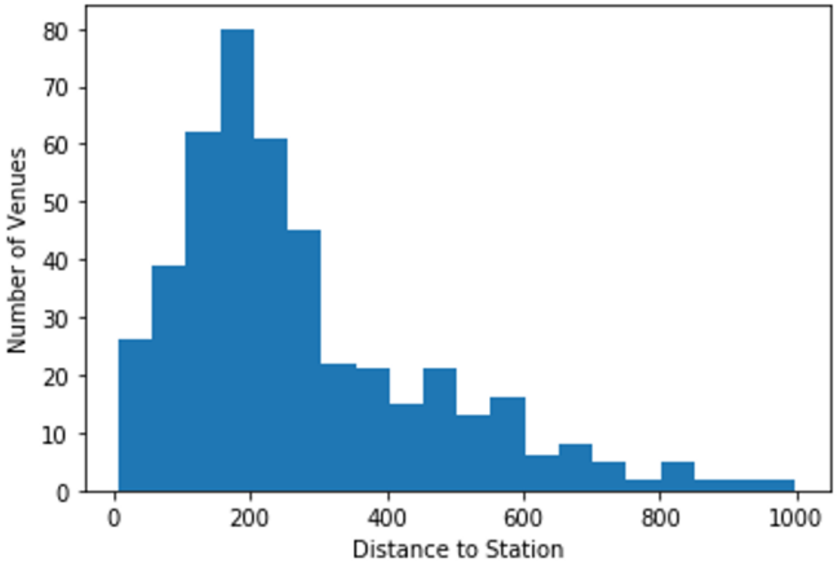

Figure 3: Histogram of Distance of Venues to Nearest Metro Station

We see that our Foursquare API calls for top venues with 1000 meters of a metro station overwhelmingly yield venues within only 300 meters of their nearest station. In contrast, if top venues were distributed evenly, we would expect this plot to grow linearly as a function of distance, due to the increasing area at increasing distance from a station. Similarly, the decrease in top venues at the shortest distances is simply due to the smaller area close to the station.

To make this notion more precise, we should examine the *density* of top venues, not the total number. In Figure 4a, we plot the density of venues (per square kilometer) as a function of distance. We see that the density falls off extremely quickly. To find the correct fitting function, it is useful to plot the logarithm of the density versus distance, which we find has a nice linear dependence, as shown in Figure 4b. The best fit line for this plot was found using LinearRegression from scikit-learn. The R-square value for this fit was determined to be about 0.97, indicating a very good fit. This means that the density of top venues actually falls off *exponentially* with distance from metro stations. As such, we recognize that proximity to metro stations is a crucial factor in the relative abundance of top-rated venues.

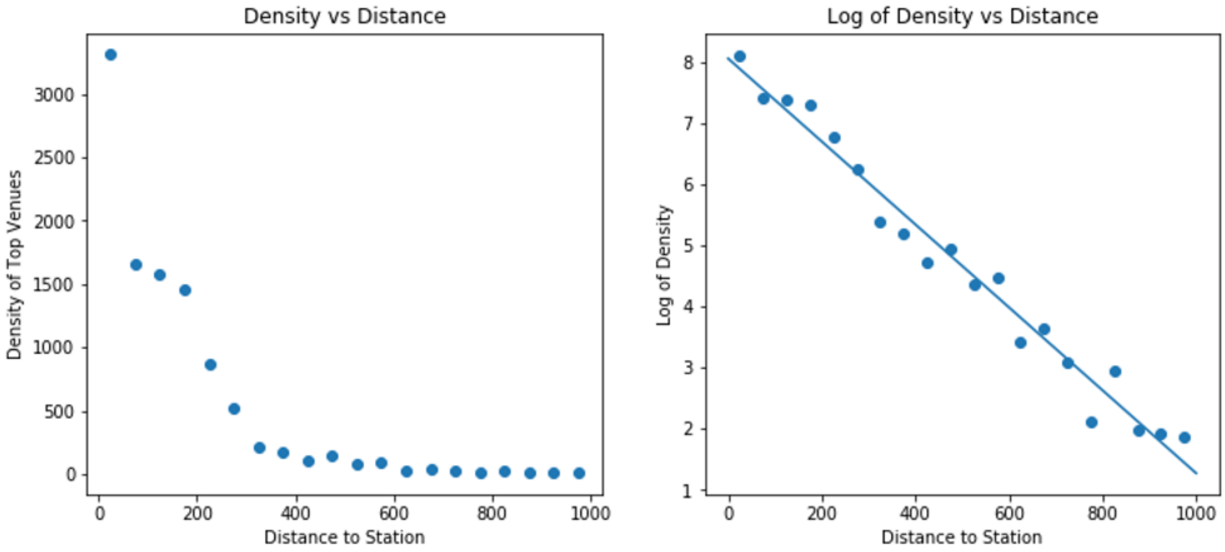

Figure 4: a) Plot of density of venues vs distance. b) Plot of the logarithm of density versus distance, displaying clear linear behavior.

We can go further still and consider how various metrics of success depend on the distance from a station. For example, a first natural guess might be to consider the ratings as a function of distance, which we plot in Figure 5.

While there is a slight downward trend in ratings as distance from a metro station increases, we see that the distribution is mostly flat. In retrospect, this should not be surprising, since our calls to Foursquare have returned only the top-rated venues, so we should expect all venues in this set to have similar ratings. If one looked at the average rating of *all* venues as a function of distance, then it seems likely that the ratings would decrease as a function of distance. However, such an analysis is not feasible with accessible data sets.

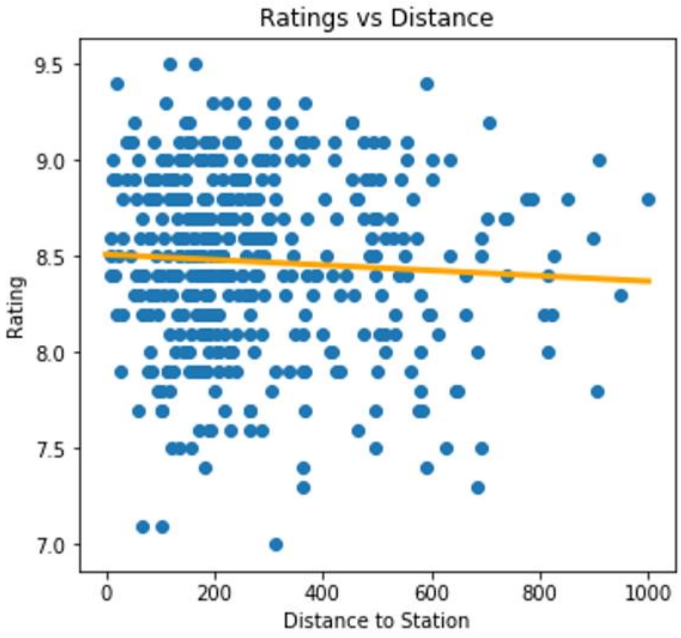

Figure 5: Plot of ratings as a function of distance, along with best fit line.

While there is little correlation between ratings and distance in our data sets, we can gain additional insight by considering the behavior of number of likes as a function of distance from a metro station, which we plot in Figure 6:

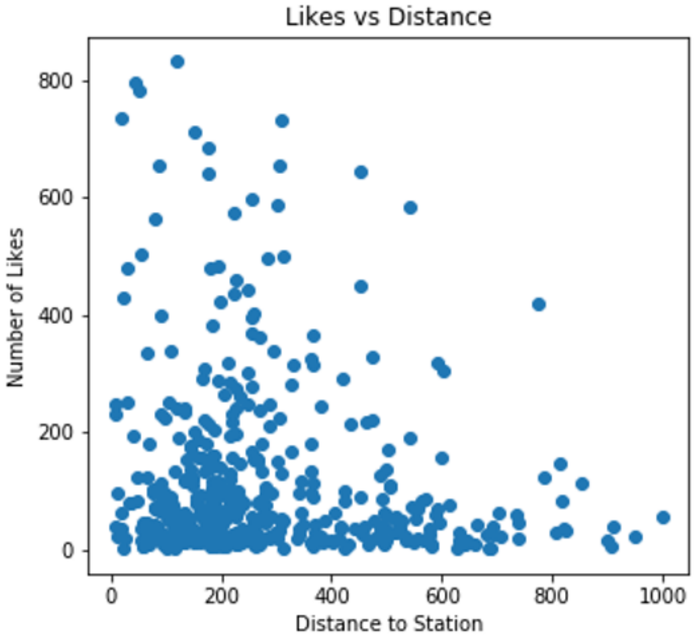

Figure 6: Number of likes vs distance

In contrast to the ratings, we see that number of likes decreases rapidly as a function of distance, indicating much higher numbers of visits to venues near metro stations, even for a fixed value of ratings. The plot does not appear to be well fit by a line. Additionally, it seems to exhibit significant heteroscedasticity, in the sense that there are much larger fluctuations in the number of likes at small distances. We can handle both of these issues by plotting the logarithm of the number of likes versus distance, as seen in Figure 7:

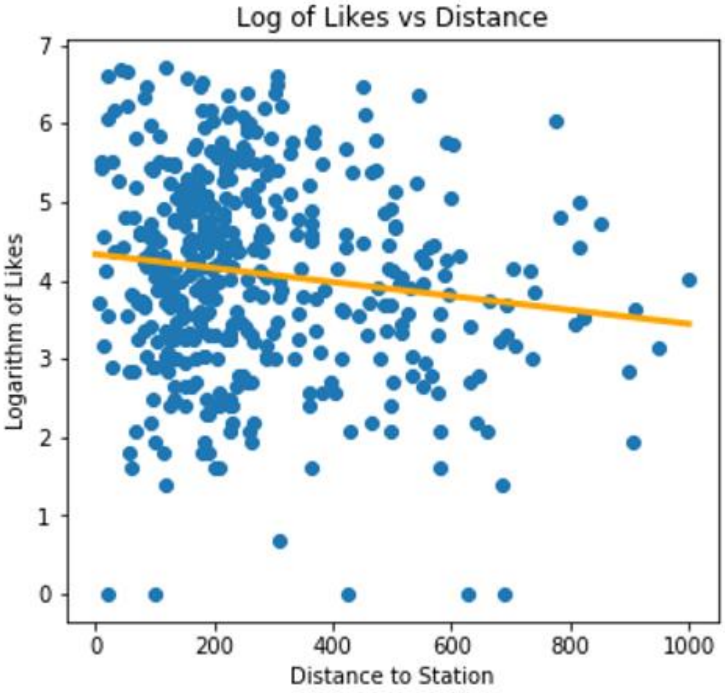

Figure 7: Log of number of likes versus distance, along with best fit line

While there are still significant fluctuations, the logarithmic plot has more uniform fluctuations, and the plot is much more consistent with a linear fit. Indeed, we find the p-value of the fit to be about 0.0046, indicating a high degree of confidence in the correlation. The roughly linear fit of the logarithm of likes indicates that the actual number of likes falls off exponentially with distance, just like the density of top venues.

Results

We have now collected various results indicating that it is generally highly beneficial for venues to be located close to public transportation, specifically metro stations. The first indication of this is that the distribution of top venues exhibits significant clustering around metro stations. Indeed, we found that the density of top venues actually decreases exponentially as a function of distance, which is a fairly dramatic trend. Note that the clustering around stations is most dramatic in the dense urban core of the city. In contrast, the map of Figure 2 indicates a few outlier stations far from the city center, where top venues are not strongly clustered around the station.

We found that ratings of top venues do not seem to vary significantly as a function of distance from a station, signaling that there actually is not much difference in quality of top venues as one moves away from a station. Nevertheless, even for venues of fixed ratings, we found that the typical number of likes for venues decreases close to exponentially as distance from the metro increases. It seems likely that this is due to the increased foot traffic present in the vicinity of metro stations, which enables more people to discover particular venues purely by chance, as opposed to by targeted advertising or word of mouth. For a typical venue, it is therefore highly advantageous to be located near to a metro station.

However, the size of this trend can vary depending on the type of venue considered. Furthermore, there are even a few outlier venue categories where the trend runs in the opposite direction. For example, in Figure 8a, we plot the logarithm of the number of likes versus distance for coffee shops, which we find have a particularly strong correlation with proximity to a metro station. In contrast, in Figure 8b, we plot the logarithm of likes versus distance for hotels, which we find actually *increase* as a function of distance, indicating that hotels fair slightly better when they are located a bit farther from a metro station. We discuss these trends further in the next section.

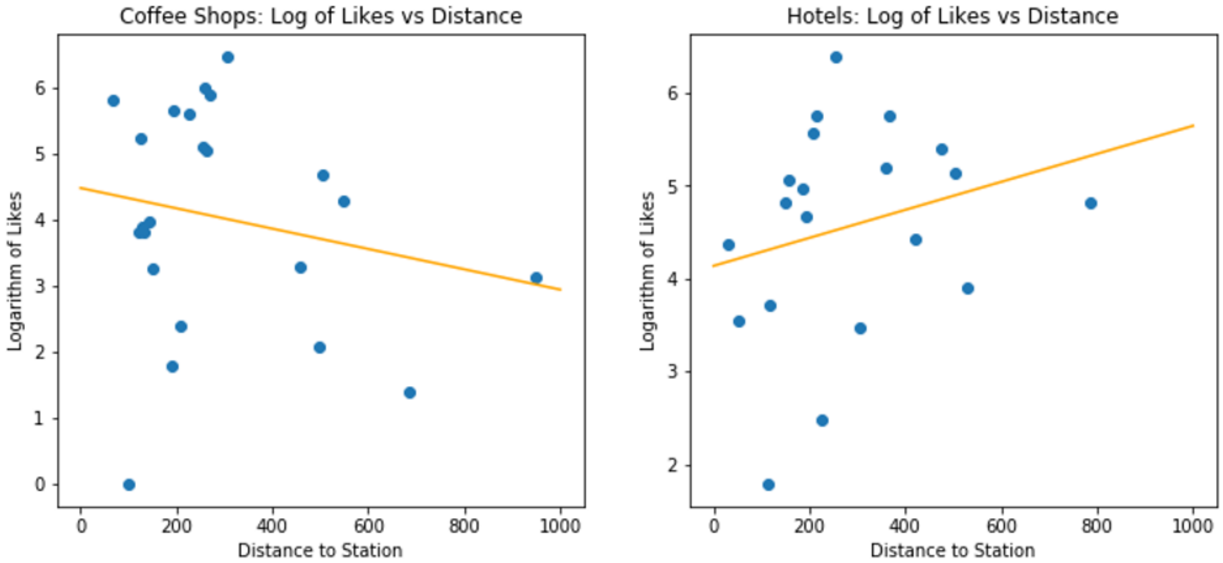

Figure 8: a) The log of likes for a coffee shop decreases sharply as a function of distance from a metro station. b) In contrast, the number of likes for hotels actually goes up as a function of distance.

Discussion

Based on the numerous results obtained in this project, we can state unequivocally that, all other factors being equal, it is typically highly advantageous for a business in a city center to be located close to a metro station. Selecting a location is crucial for prospective new business owners, and the exponential decrease of likes indicates that distance from the metro is one of the most important factors in determining the ideal location.

For a randomly chosen type of new business, it is almost certainly true that locating near a metro station is the ideal choice. For certain types of venues, such as coffee shops, the amplification of business due to metro proximity is particularly pronounced. (Presumably, this has to do with the fact that metro commuters want to get coffee on their way to work.) However, there are certain exceptions to be borne in mind. For example, we saw that hotels actually seem to perform slightly better when they are slightly farther from a metro. Perhaps this indicates that hotel guests prefer to have a slightly quieter neighborhood, away from the crowds and noise of metro station, but this issue would require further study.

As such, for a prospective new business owner, we would encourage them to look at the trend of likes versus distance for the particular type of venue that they wish to open, before committing to opening as close to a metro station as possible. However, in the absence of sufficient data for a particular type of venue, it seems fairly safe to recommend that a generic business type should try to open close to a metro station.

Conclusion and Outlook

In this project, I have shown that business venues located near a metro station do significantly better than their counterparts farther from public transportation. Even when considering businesses receiving the same ratings, venues farther from a metro station receive a much smaller number of likes (decreasing exponentially with distance), indicating a much smaller number of visits from potential customers. The exponential nature of this trend indicates that proximity to public transportation is a crucial factor in picking the location of a new business.

However, there are certainly other factors to consider in picking the location of a business. For example, presumably the rent associated with locations close to metro stations is higher than the rent for buildings farther away. Furthermore, the large number of venues close to a metro station certainly results in an increased level of competition for these venues. There is most likely a saturation point, i.e. a critical density of venues, above which venues close to a station will start to perform worse than distant venues. A future analysis could take these factors into account, for example by finding concrete data on the rent and revenue of businesses.

It would also be interesting to see how well these results generalize to other cities, as well as to other modes of public transportation. It seems natural that similar correlations will be found for other large cities with well-developed subway systems, such as New York City and Washington D.C. However, it is less clear that any significant correlation will be found in smaller cities where buses are the primary means of public transportation. In smaller cities, a much larger fraction of the population gets around via driving a personal automobile, which will certainly weaken the trends observed for larger cities.