**IBM Data Science Certificate Capstone Report**

**Jordan Lieberman – June 2020**

I. Introduction

The city of Boston, Massachusetts and its' surrounding areas (ex. Cambridge) is a thriving city on the east coast of the U.S. Many industries, especially healthcare and biopharma have placed their headquarters in the region. The region is also home to dozens of colleges and universities that bring tens of thousands of students to the area. Unfortunately, this economic activity comes with a cost, congestion. The Boston metro area is notorious for its traffic. Depending on the study, it has been ranked as the worst traffic in the entire United States. In fact, the Boston Globe published a wide-ranging analysis of the problem in November 2019 called, "Seeing Red" that documented the problem and some potential solutions.

One piece of the solution to this problem is for more people to use public transportation such as trains and buses. In the Boston area, one key method of public transit is the MBTA, or Massachusetts Bay Transportation Authority. The MBTA offers subways, ferries, commuter rail, and buses as means of moving about the area. In this project, we'd like to focus specifically on the subway. We would like to know what types of businesses (and how many) are near each subway station. This analysis will give us an idea of which stations might be best to live and/or work near.

Possible audiences for our analysis include people moving to Boston from other areas, those moving within the city, and perhaps, most importantly, those considering getting rid of their car in favor of commuting on public transit. We want to help these individuals decide which subway stations might be best suited to live near. The analysis could also be useful for businesses in deciding which stations to locate their offices near. A key benefit for many employees is the ability to go the gym after work, grab a drink with colleagues, or have lots of lunch options within walking distance.

II. Data

To complete this analysis, we will use geographic data for each subway station as well as the FourSquare API to explore venue information within a specified radius of each station. We will identify the latitude and longitude of each MBTA subway station in the region. Then we will use the FourSquare API to pull the types and number of venues within a specified radius of each station.

First, we used the MBTA’s [Developer Page](https://www.mbta.com/developers/gtfs) to download the latitude and longitude of each station in the Boston metro region. This data was reasonably well accessible and easy to understand and use. It required minimal cleaning and organizing to get the most important features for our analysis.

We then used each set of latitudes and longitudes as input to the FourSquare location API. These coordinates coupled with a radius of approximately 1/3 mile (500 meters) and a limit of 40 on the number of venues allowed us to complete our table of venues for each MBTA station.

III. Methodology

In order to have a useful analysis, we had to do a lot of cleaning of the data. First, with the MBTA station data we needed to drop most of the attributes, such as ‘vehicle type’ and ‘parent station’ which were not useful for our project and would only slow down processing of the data. Eliminating extraneous data also helps to prevent user confusion when viewing the final table.

Next, we needed to filter the stops and create a new table for only rapid transit (i.e. subway). This was accomplished using Boolean logic and the ‘zone id’ attribute. We then needed to eliminate the duplicates from the list. The duplicates occur because the MBTA data includes one row for each platform. In other words, for a north-south train, there is one row for the northbound platform and one for the southbound platform. Since these trains use the same station location, we remove these to avoid confusion and make the analysis faster.

The next piece of our analysis required putting together a map using the folium library. We first created a blank map of the Boston region. We then used the coordinates of each MBTA station to add markers to the map so that the user can easily see where the stations are and how far they reach into the region. This is especially helpful in understanding how the lines relate to one another and where the subways are in relation to each city and town as well as major highways and thoroughfares.

We also added color coding to the folium map markers so that users can differentiate between the different lines of the subway.

The next step in the analysis was to use the FourSquare API to identify venues that are in close proximity to each MBTA rapid transit station. We used a radius of 500 meters (approximately 1/3 mile) as a good estimate of a short walk from each station. This is representative of running out to grab lunch during work or to a nearby bar after work with colleagues. We also limited the total number of venues to 40 so that we did not overload the dataset with results. This was particularly important for the stations in downtown because we anticipated that there would have been far too many results to effectively manage.

The next task was the most time-consuming aspect. The FourSquare results tag each venue with a category. However, these categories are too descriptive and nuanced to do any real analysis. There were approximately 250 different venue categories which made any kind of comparison very difficult. To solve these, I created 5 generic groups of venues to place each venue in. These were ‘food and drink’, ‘retail’, ‘services’, ‘exercise/outdoors’, and ‘nightlife’. These categories represent main types of venues near each station that users might be interested in.

I also removed a variety of venue types that were duplicative or likely would not be of interest to users. For example, ‘metro station’ was removed because these stations were the center of our analysis and thus duplicative. I also removed items such as ‘college rec center’ and ‘border crossing’ as these did not seem relevant.

The last, but most important, step was to pivot the dataframe so that we could count the number of each venue type within the specified proximity of each station. Once we had this table, we added a total/sum column and replaced all NaN values with zero so that the totals would calculate correctly. Finally, we sorted the table on the total column so that we can easily see which MBTA stations have the most venues near them.

As a final step, we added markers for all the venues to the folium map as a visual aid for the user.

IV. Results

There are 118 rapid transit stations in the Boston metro area. These stations are part 4 different subway lines that cross the region: red, blue, green, and orange. Of these 118 stations, 9 had the maximum number of venues possible (40) within a 500-meter radius. This provides a good indication of how busy a particular area might be and how desirable it could be to live or work nearby.

The next group of stations had between 30 and 39 venues within a 500-meter radius. There were 37 stations in this category. These represent busy areas that would also be desirable to live or work near.

Next, there were 24 stations with between 20 and 29 venues within a 500-meter radius. These are moderately busy areas that might still be desirable areas but also represent an opportunity for quieter locations that still maintain access to the MBTA system for quick transportation to other areas of the region.

In the next broad grouping, there were 30 stations with between 10 and 19 venues inside our 500-meter radius. These stations occupy significantly less dense areas than our previously discussed categories. Again, these areas may still be desirable because they are connected to the rest of the region by the rapid transit system.

The remaining 18 stations at the bottom of our list have less than 10 venues within a 500-meter radius. These likely represent the least desirable areas to live or work in terms of access to services, retail, and dining options.

V. Discussion

One surprising aspect of this analysis is that several stations at the top of our list are outside the downtown or central city location. This could be very useful information as it is common to quickly focus on the center of the city.

For example, stations such as Copley, Tufts Medical Center, and South Station are to be expected at the top as these are very busy, well developed areas in the center of Boston. On the other hand, Griggs Street and Newton Centre are much further from downtown and might surprise a user that is familiar with the city. These areas represent a potential opportunity.

For someone who is looking to move in or within the region, stations such as Griggs Street may provide an opportunity for a lower cost compared to the center of downtown. They might also represent a moderately quieter area compared to the much denser areas in the financial district.

For a business that is looking to locate their offices in the region, these same areas might represent a lower cost for commercial real estate. They might also represent areas that have been underdeveloped and could be in cities and towns that are anxious for more commercial tenants and the taxes they bring along.

In both cases, identifying areas outside the city that remain desirable due to their abundance of retail and nightlife options present particularly good prospects for reducing the number of people (and cars) entering the city center. This would reduce congestion on the highways, enabling people to move throughout the city more efficiently. There has also been a variety of evidence that people are happier and more productive when they don’t lose significant portions of their day to traffic.

Of course, this analysis is ripe for expansion should we want to incorporate other data sets. Other factors contributing to a desirable area might include school quality, crime data, or proximity to other transportation options such as airports, commuter rail, and buses. For businesses, we might also want to look at the cost of utilities (e.g. water) or the tax base to identify areas ripe for development at a lower cost.

VI. Conclusion

This analysis provides a good opportunity to take a different look at the Boston metro region. Using data from the Massachusetts Bay Transportation Authority (MBTA), FourSquare, as well as geographic coordinates from Python libraries, we were able to identify rapid transit stations based on their density of desirable venues.

After substantial cleaning and organizing the data, we were able to categorize and sort the MBTA stations based on the type of venues nearby. We also provided a visual aid using the folium map library to help the user better understand the region.

Overall, we were able to group stations into 5 categories based on the total number of venues. Users could also sort based on the more granular category, such as nightlife, if they wanted to see a different ranking of the stations. This provides a good overview of the Boston metropolitan region through the lens of subway stations and the associated businesses within a close radius of each station.