Capstone Project - Proposing Locations for a Bubble Tea Store in Vancouver (Week 2)
Applied Data Science Capstone by IBM/Coursera

Introduction/Business Problem

This project proposes good locations in Vancouver for opening a bubble tea store. It is aimed to provide help to any person wanting to start a business in Vancouver with relatively small initial capital or a chain that wants to open a location in Vancouver.

Bubble tea is a drink that can be made using a variety of ingredients such as tea, juice or sweetened flavoured milk usually with sweet black tapioca pearls.

Figure 1: Bubble tea(image from https://www.thecoast.ca

Starting a bubble tea store is relatively inexpensive, since it requires less equipment than a traditional coffee shop or restaurant and the profit margins are high because the ingredients, such as tea, are inexpensive. Also, preparing bubble tea is easy, so not much staff training is required.

Bubble tea originated in Taiwan in the early 1980s and is gaining popularity around the world. For example, the Taiwanese chain Chatime currently has 2500+ locations around the world (https://en.wikipedia.org/wiki/Chatime) but only 3 in Vancouver.

The number of all bubble tea stores in Vancouver is around 60 (results from https://foursquare.com), whereas there are 108 Starbucks in Vancouver as of February 2019 (https://www.statista.com/statistics/306896/cities-with-the-largest-number-of-starbucks-stores-worldwide/) and there are also numerous other coffee shops.

Bubble tea is especially popular with tweens and going for bubble tea after school is loved by Vancouver school children.

Because of this, to propose a good location for a bubble tea store we look for locations in Vancouver that are close to schools. In particular, we look at schools that have no bubble

tea stores within 800m, determine if a school is "walkable" and highlight the commercial zones that allow for opening of a bubble tea store near the school.

We say that a school is "walkable" using a definition of a walk score (a number between 0 and 100) on https://www.walkscore.com/methodology.shtml:

> *Walk Score measures the walkability of any address using a patented system. For each address, Walk Score analyzes hundreds of walking routes to nearby amenities. Points are awarded based on the distance to amenities in each category. Amenities within a 5 minute walk (.25 miles) are given maximum points. A decay function is used to give points to more distant amenities, with no points given after a 30 minute walk.*
>
> *Walk Score also measures pedestrian friendliness by analyzing population density and road metrics such as block length and intersection density. Data sources include Google, Education.com, Open Street Map, the U.S. Census, Localeze, and places added by the Walk Score user community.*

The commercial zones we focus on have high pedestrian traffic, which is also very good for a bubble tea store.

Data

To recommend good locations for a bubble tea stores as outlined in the Introduction we need the following data:

- **The number and locations of all bubble tea stores in Vancouver,**
  These are obtained from https://foursquare.com/ The names and the latitude and longitude of all bubble tea stores can be obtained from here.

- **The locations of all elementary and high schools in Vancouver,**
  These are obtained from https://foursquare.com/ The file includes the names, addresses,categories and longitude and latitude of the schools. This is important for both data processing and visualization.

- **The locations of all commercial zones.**
  These are obtained from https://data.vancouver.ca/datacatalogue/zoning.htm. I will use the .kml file to extract the polygons that choropleth maps need, so I can then visualize the zones on a map. The .kml file is converted into .geojson file using the online converter at: https://mygeodata.cloud/converter/kml-to-json

- **The walkability of school locations,**
  These are obtained from https://www.walkscore.com using their API.

Methodology

According to https://foursquare.com/ , there are 60 bubble tea stores in Vancouver. The names, latitudes, longitudes and addresses of the stores are loaded into a data frame `bubble_tea_df`. Minor data cleaning is required as not all stores have the address specified. However, all stores have the latitude and longitude specified.

According to https://data.vancouver.ca/datacatalogue/schools.htm there are 194 schools in Vancouver. The content of the `csv` file that can be downloaded from that website is loaded into the data frame `schools1`. No data cleaning is necessary.

There are 180 zones that allow for commercial development in Vancouver. We are only looking at "C" or "FC" zones. We are not including the Comprehensive Development District (CD) zones listed at https://vancouver.ca/home-property-development/cd-1-comprehensive-development-district.aspx as these zones are either very residential (no possibility for opening a bubble tea store) or very commercial (likely already contain a bubble tea store). The zoning data downloaded from the website is in .kml format. An online converter at https://mygeodata.cloud/converter/kml-to-json was used to convert the data to .geojson form and than the required zones were extracted to a data frame `zoning`.

Before doing the analysis we visualize all the data.

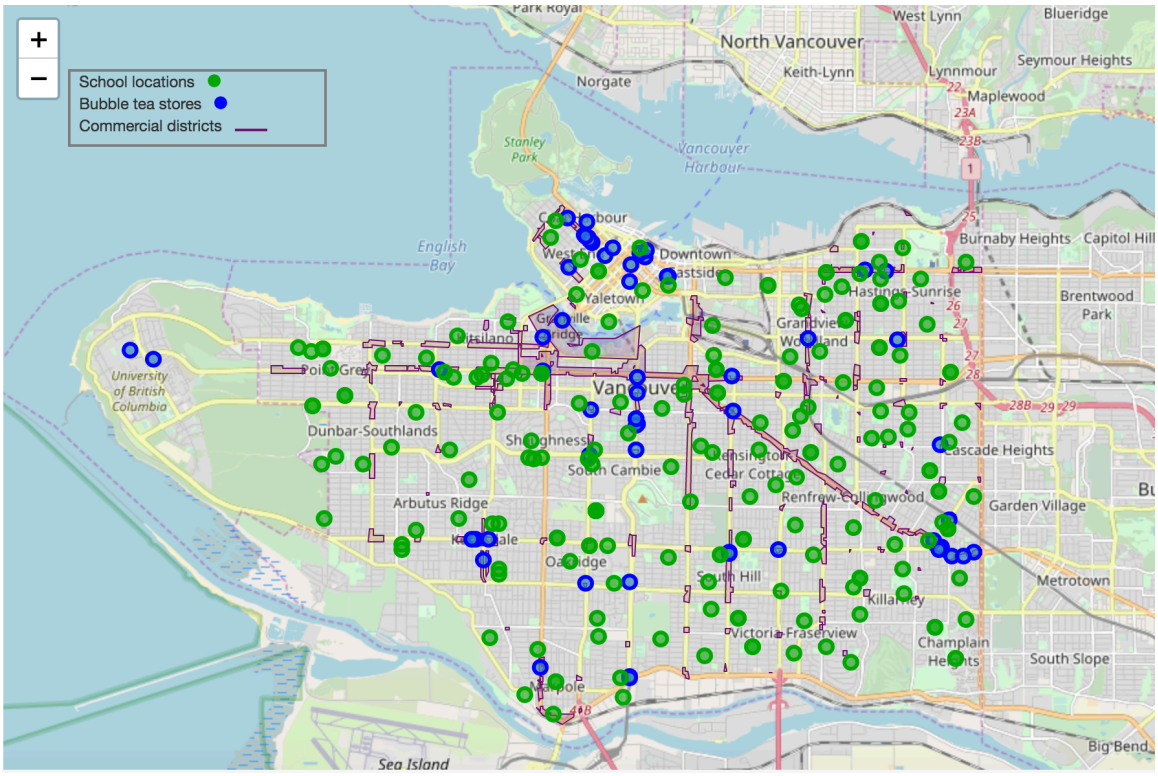

Figure 2: Location of bubble tea stores, schools and commercial zones in Vancouver

We can notice that the schools are evenly spaced around the city, but there seem to be clusters of bubble tea stores.

Clustering schools to find the best location for a bubble tea store might not work, because it could produce a location that is way outside of any commercial zones.

Instead we look for locations that are in commercial zones, within 800m of a school which has no other bubble tea store within 800m and is in a very walkable area.

The easiest way to look for locations is to look at all the schools in the city and eliminate the ones that have a bubble tea location within 800m and also eliminate the ones that have no commercial districts within 800 m.

We choose count the bubble tea stores that are within 800m from a school. The value of 800m was chosen, because it can be walked in about 10 minutes, which seems like a reasonable distance to walk for school aged children.

We only look at schools that are at most 800m away from a commercial district that allows for opening of a bubble tea store. The commercial zones are specified by the polygon of longitudes and latitudes (from the geojson file), and the distance between a school and a commercial zone will be the minimum distance between the school and a point in the polygon.

It is important to notice that the above distances of 800m and 0.8km are coded as constants, so if the stakeholders would like to consider different values, the analysis could be easily redone.

To enable the data processing, first we add columns to the `schools1` data frame. For each school, using the `zoning` and `bubble_tea_df` we add the information, so that the data frame contains the following columns:

SCHOOL_NAME – the name of the school
LATITUDE – the latitude of the school
LONGITUDE – the longitude of the school
ADDRESS – the address of the school
WALK_SCORE – walk score of the school (retrieved from https://www.walkscore.com using the latitude, longitude and the address of the school)
BIKE_SCORE –bike score of the school (retrieved from https://www.walkscore.com using the latitude, longitude and the address of the school. Not used in this analysis, but could be useful for future analysis)
NUM_CLOSE_STORES – how many bubble tea stores are within 800m of the school (retrieved from https://foursquare.com/)
CLOSE_DISTRICTS – list of indices of districts which are within 800m of the school (calculated using the above specified method)
NUM_DISTRICTS – length of the list of CLOSE_DISTRICTS

Now we have the all that is needed for the analysis.

The walk score measure seems to be a much more sophisticated measure of pedestrian friendliness then the number of near by commercial districts. In fact, the map at http://vanmapp.vancouver.ca/pubvanmap_net/default.aspx?&showLayers=zoning_district_labels,zoning_districts shows that in some areas there are quite a few small commercial districts, i.e. that commercial districts in Vancouver can be quite fragmented.

We calculated the Pearson correlation coefficient and only 25% of the walk score variations can be explained by the number of nearby commercial zones

Even if we take into account that the commercial districts can be fragmented and use a binary variable for each school that denotes if a school is close to a commercial district or not, and then use Point-Biserial Correlation Coefficient that measures the strength of association between a continuous-level variable and a binary variable, we don't get a much better explanation of the variation.

However, there are no schools with many commercial districts nearby that have a low walk score.

Therefore we use the walk score in our analysis to rank the schools that don't have any close by bubble tea stores, but we do eliminate the schools that do not have a close by commercial district from our analysis.

There are only 10 schools in Vancouver that do not have a close by commercial district. All but two of these stores are in very residential neighbourhoods. Alexander Academy is in CD zone and Royal Canadian College has a walk score of only 80, which is, for Vancouver, relatively low.

The schools that have no nearby bubble tea stores, but do have a near by commercial zone are sorted by their walk scores in descending order and we further examine the top 5 schools in this ranking.

In order to produce a good visualization map, we colour the commercial zones that are close to our top 5 schools and the top 5 schools differently.

Here is a map on which these 5 schools and the nearby commercial districts are coloured red.

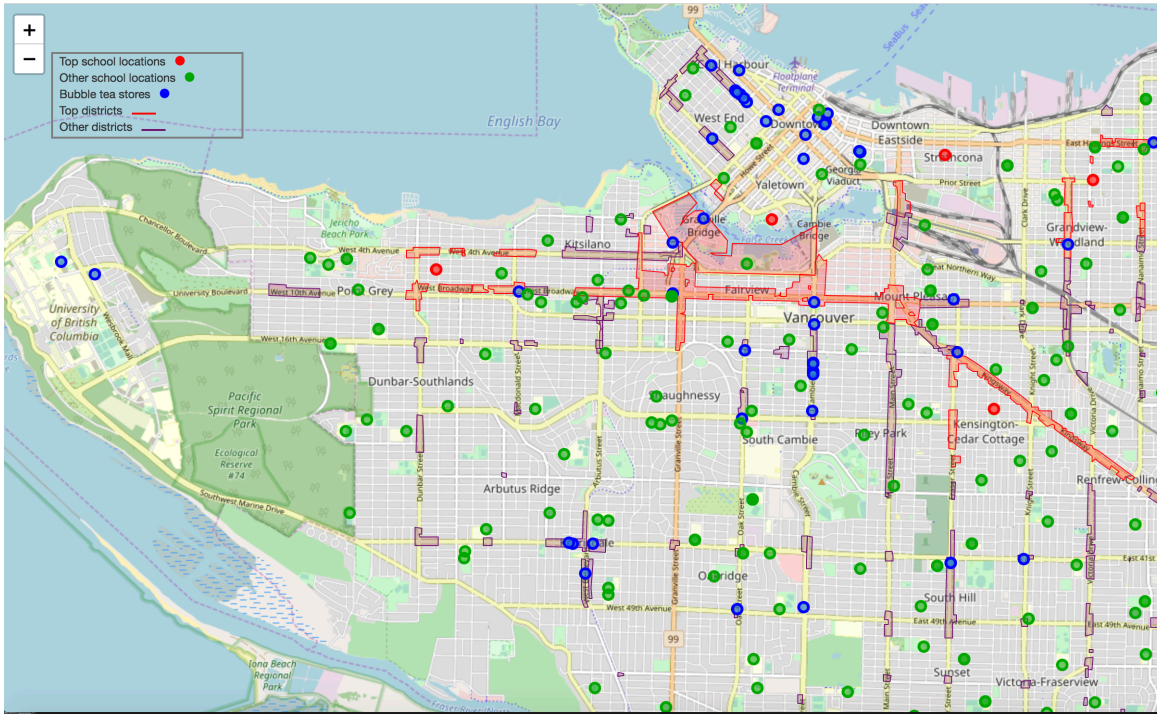

Figure 3: Recommended locations for bubble tea stores

The top 5 schools are:

1. Elsie Roy Elementary
2. Lord Strathcona Community Elementary
3. Bayview Community Elementary
4. St Francis of Assisi
5. Charles Dickens Annex

Results

Our program identified five schools in Vancouver that do not have a bubble tea store in their vicinity, but are close to a commercial district, so it is possible to establish a new store.

The map we created is also a very good visual tool for exploring where the new store should be located.

It is interesting that most bubble tea stores are located in a commercial district we identified. Others are located in CD districts that we chose not to examine (such as downtown), and from the map one can see that there are already many bubble tea stores there.

The Elsie Roy Elementary school was recommended based on the proximity to the zone that is across the water from it, and as such is really not at a walking distance to the zone. One could eliminate such situations by refining the search algorithm, but this example also shows how important is to have a human interpret the results.

The top 5 ranked schools are have very similar walk scores and it would take an expert on the area to distinguish between them. Based on other factors, such as the size of the schools, how many other schools are relatively close to the chosen school and the make up of the neigbourhood, I would recommend opening a bubble tea store close to a schools using this ranking:

1. Charles Dickens Annex
2. Bayview Community Elementary,
3. Lord Strathcona Community Elementary (consider the historic Chinatown district).

Discussion

This project brings up two very important points when it comes to machine learning. The first one is that unbiased algorithms can find good locations for starting a new business. For example one of the schools, Charles Dickens Annex, is quite far from the city centre and might have been overlooked by potential stakeholders, but is a very good location for opening a bubble tea store.

The second point is that human knowledge is still required to assess the output of any algorithm. In our project, one of the schools chosen was Elsie Roy Elementary that is close to a commercial zone that is across the water, and as such not really easily walked to.

Conclusion

The purpose of this project was to identify a good location for a bubble tea store in Vancouver. Since bubble tea is very popular with tweens, we decided to focus on locations that are close to a school. The algorithm we created chose 5 schools that have a high walk score and no close bubble tea stores, but that are close to a commercial district, so opening a bubble tea store in the vicinity is possible.

Furthermore, using human knowledge of the area, the 5 locations are narrowed down to 3.

Final decision on the optimal location for a bubble tea store will be left to the potential owner who might also need to consider other factors, such as availability and cost of renting a location.