Capstone Project - The Battle of Neighborhoods

27-05-2021

Introduction

Background

Since its inception in 1984 the gastropub has become a true staple in most major cities. A weary passer-by may stop by any such establishment and be treated to a variety of inventive food and drinks. It could very well be argued that the gastropub is one of the major bringers of life to cities today.

As such, there are many businessmen eager to open such an establishment. Among them is our client, a restaurateur from London looking to make a name for himself. While his business is doing well enough in his hometown he is looking to expand into Manchester. Because our client is not familiar with the area, he has hired us to gather insights into a good location for a new gastropub.

Problem

Using data we will be able to determine which neighborhoods in Manchester have a large or low amount of gastropubs. We aim to find the areas with the lowest amount of gastropubs, so that the client knows where he can open his new establishment.

Interest

Obviously, our client is most interested in an accurate depiction of which neighborhoods have the lowest frequency of gastropubs. Those who are interested in moving to an area of Manchester with a high (or low) frequency of gastropubs may also be interested in this information, as well as other restaurateurs and pub owners.

Data

We will be using the Wikipedia page which lists all areas in Manchester (https://en.wikipedia.org/wiki/Category:Areas_of_Manchester) to gather a list of neighborhoods in Manchester. We will filter any wards out of this list as they overlap with neighborhoods and as such are not interesting to us (we already have a smaller grain of the data).

Using this list of neighborhoods, we will gather a list of latitudes and longitudes using the ArcGIS geolocator. We are using this geolocator as it is free to use and reasonably accurate. We specify that we only want coordinates located in Manchester, UK, so we can be sure that ArcGIS will not give us any incorrect locations. After drawing the coordinates on a map it is clear that no coordinates lie outside of Manchester.

We will also use the Foursquare API to gather information about establishments located near every neighborhood. We will get the top 100 venues for each neighborhood within a 2km radius. We only take the top 100 to keep the tool manageable but also to make sure that we only get relevant results. We are not interested in gastropubs that are only visited by a very low amount of people or has a very low appreciation by the general public as this does not form any real competition to our client.

Methodology

Our methodology is rather simple. After gathering the data we group rows by neighborhood and use one hot encoding to make a row for every type of venue. We then take the mean of the frequency of occurrence of each category of venue. This gives us a normalized number between 0 and 1 which we can use to determine the frequency of each type of venue per neighborhood.

Because we are only interested in gastropubs we filter out any data relating to other venues before clustering the data.

For this, we use k-means clustering based on the frequency of gastropubs in every neighborhood. Doing so will place each neighborhood within one of three clusters.

We use k-means clustering because we know the amount of clusters we want beforehand. We want to have three clusters, for a low, medium, and high amount of gastropubs. We believe that this makes the data more comprehensive to the stake holder because the amount of categories is limited which makes making a decision simpler. In other words, three categories are easy to understand while more than that would be unwieldy and less would give an oversimplified image.

We also don't care about excluding any "outliers" as all our data is normalized and between 0 and 1. As such, we do not believe that there are any extreme outliers that will have a negative impact on the clustering (noise) – which leads us to k-means clustering instead of DBSCAN.

Results

We have placed every neighborhood in one of the following categories:

- High
- Medium
- Low

Our most important result is a map which shows every neighborhood categorized into one of the three clusters. We also have created a table for every cluster. This gives the stake holder a comprehensive view of the data as he can clearly see which of the three cluster every neighborhood falls and which part of town is more likely to be in one cluster or the other.

The map is included in this repository as the file map_clusters.html.

Discussion

It is very clear that the south of Manchester is quite busy when it comes to gastropubs. As such, we do not recommend the client to open an establishment here.

The west of Manchester also has a reasonable amount of gastropubs, but seems to be less "saturated" than the south.

We have discovered, however, that the east of Manchester seems to be almost devoid of gastropubs. We believe this would make ideal location for our client to open an establishment.

Conclusion

Using the data from a multitude of sources we have been able to create a clear view of the frequency of gastropubs in every neighborhood in Manchester. We believe that based on the information provided by us the client is in a better position to make a decision on where to open his establishment. Mainly, we believe that the east of Manchester would make an ideal location as it is almost devoid of gastropubs.

If this is not a possibility for any reason (e.g. scarcity of properties in this region) we recommend the customer to open an establishment on the west side of Manchester, as the number of gastropubs is relatively low here as well.

Thank you for reading my report.