Applied Data Science Capstone Project
- The Battle of Neighborhoods

Anthony Swain, PhD

February 23, 2020

1. Introduction

1.1 Forward

Many of you who are reading this will already be aware of the context of this post. However, for those who are not aware, this report is written in connection with the Applied Data Science Capstone Project for the IBM Data Science Professional Certificate on Coursera [1]. Students of this course are spending the first 8 modules focusing on building up data science skills with the course culminating in a large capstone project where the students are able to exhibit some of the skills they learned throughout the course. It is essential for us to use Foursquare location data in this final project and come up with a creative business case to exhibit our skills.

1.2 Background

I have been working as an expat in Asia since 2006. My background is in molecular biology and as such I have not had much of an opportunity to work on location data before. I was born and raised in the UK, my wife is from Thailand and I live in Japan. Having this mixed background gives me a lot of potential locations to work with, however based on the potential of multiple non-English languages encroaching on this project, I decided to pick my hometown; Leeds, UK, as the topic of this project.

Leeds is a city in the North of England located about 170 miles North of Central London [2]. According to a 2011 census, the population of Leeds is around 2.6 million [3] and is likely to be on the rise. I did my undergraduate degree at one of the five universities in Leeds [4] and found the city to be a very active town that is popular amongst students. As a student I my friends and I would often try to visit different bars and restaurants in the area, but hindsight has led me to realize that there may have been something missing in the city…Thai cuisine. Which brings me to the business problem.

1.3 Business Problem

In the event that I move back to the UK I would want to move back to my hometown with my family. In such a scenario, it would be feasible for me to open a business in the form of a restaurant. My wife is an amazing cook and I think her talent for cooking could really bring in repeat customers. Although there were almost no Thai restaurants in Leeds when I left in 2006, the landscape may have changed a lot. This data science project should be able to 1) Give an overview of the current restaurant landscape in Leeds, 2) Understand the current competition with regards to Thai restaurants in Leeds, 3) Pick out some possibilities for restaurant locations with minimum competition and an appropriate customer base.

1.4 Interest

This data science project should be interesting for anyone who is interested in the current restaurant landscape of Leeds. It could also be of interest to anyone who is realistically interested in opening up a Thai restaurant in Leeds (in a non-hypothetical sense). Also, given the structure of how this report is written, it may be of interest to anyone who wants to perform a similar analysis in other cities in the UK or around the world.

2. Data

2.1 Data Sources

In this project there are three main sources of data used. The first source of is the Wikipedia page for the Leeds Postcode area [5]. Wikipedia pages are not generally sources that I would use academically but the Wikipedia page uses the Office of National Statistics [6] to get this information (the presentation of the data on in the Wikipedia source is in an accessible format which is appropriate for this course). The Information obtained from the Wikipedia page will be used as the base for the data frame using the postcode district and area (coverage) as the main points. The next source of data is from the website freemaptools.com which has a .csv file directory of all of the postcodes in the UK and their respective latitudes and longitudes [7]. This data is used to complete a base data frame to connect to the post codes and area names of the initial data frame based on the Wikipedia table. The final main data source is Foursquare [8]. Foursquare is an application we have been using through this course and is a technology company which built a massive library of location data. The Foursquare library is used in conjunction with the location data I got from the first two sources in order to perform data analysis of the venues in the Leeds area.

References

1. Coursera Course - https://www.coursera.org/professional-certificates/ibm-data-science
2. Leeds definition - https://www.collinsdictionary.com/dictionary/english/leeds
3. Office for National Statistics - http://www.nomisweb.co.uk/articles/747.aspx
4. Leeds Wikipedia - https://en.wikipedia.org/wiki/Leeds
5. Leeds LS Wikipedia Page - https://en.wikipedia.org/wiki/LS_postcode_area
6. Office for National Statistics Source Data - www.ons.gov.uk/ons/guide-method/geography/products/postcode-directories/-nspp-/onspd-user-guide-and-version-notes.zip
7. Freemaptools.com postcode data - https://www.freemaptools.com/download/outcode-postcodes/postcode-outcodes.csv
8. Foursquare - https://foursquare.com/