Cover Page

Capstone Project - The Battle of Neighbourhoods: Welsh Towns

Applied Data Science Capstone
by IBM

Part of our IBM Data Science
Professional Certificate

Author: Rafal Radecki

Version: 01

Date: 15/05/2021

Instructions

Now that you have been equipped with the skills and the tools to use location data to explore a geographical location, over the course of two weeks, you will have the opportunity to be as creative as you want and come up with an idea to leverage the Foursquare location data to explore or compare neighborhoods or cities of your choice or to come up with a problem that you can use the Foursquare location data to solve. If you cannot think of an idea or a problem, here are some ideas to get you started:

- In Module 3, we explored New York City and the city of Toronto and segmented and clustered their neighborhoods. Both cities are very diverse and are the financial capitals of their respective countries. One interesting idea would be to compare the neighborhoods of the two cities and determine how similar or dissimilar they are. Is New York City more like Toronto or Paris or some other multicultural city? I will leave it to you to refine this idea.
- In a city of your choice, if someone is looking to open a restaurant, where would you recommend that they open it? Similarly, if a contractor is trying to start their own business, where would you recommend that they setup their office?

These are just a couple of many ideas and problems that can be solved using location data in addition to other datasets. No matter what you decide to do, make sure to provide sufficient justification of why you think what you want to do or solve is important and why would a client or a group of people be interested in your project.

Review criteria

This capstone project will be graded by your peers. This capstone project is worth 70% of your total grade. The project will be completed over the course of 2 weeks. Week 1 submissions will be worth 30% whereas week 2 submissions will be worth 40% of your total grade.

For this week, you will required to submit the following:

- A description of the problem and a discussion of the background. (15 marks)
- A description of the data and how it will be used to solve the problem. (15 marks)

For the second week, the final deliverables of the project will be:

- A link to your Notebook on your Github repository, showing your code. (15 marks)

2. A full report consisting of all of the following components (15 marks):

- Introduction where you discuss the business problem and who would be interested in this project.
- Data where you describe the data that will be used to solve the problem and the source of the data.
- Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.
- Results section where you discuss the results.
- Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.
- Conclusion section where you conclude the report.

3. Your choice of a presentation or blogpost. (10 marks)

Data

The first task is to obtain a list of towns or localities in Wales. Wikipedia holds a list of 446 localities with their population in a single table. This source is reliable, will be easy to scrap, and filter appropriate size localities. Because the couple is looking for a quiet place yet with some vibrant community, hence it should not be a small village, nor a large town. This report focuses on localities with population raging from 2,000 to 20,000 citizens.

Four Square database contains very comprehensive information about various type of venues, which could be used to identify similar towns based on composition of similar businesses.

Information about school performance could be downloaded from the internet as a structured dataset. All state schools in Wales are graded using a colour code, where excellent schools are 'Green/Gwyrdd' and good schools are 'Yellow/Melyn'. Although parents should read detailed information about the school and not focus only on the colour code, this goes beyond the scope of this project.

This report uses the latest secondary school database from year 2019, because of assumed age of children.

UK government website contains a structured data set of average house prices in xlsx format. This report uses the average price for a detached house per county in Wales on 01/04/2020 (the latest available). The data set contains historical data for the whole of UK, hence it has to be filtered using the most recent date and the list of counties in Wales.

List of resources:

- Localities with population: https://en.wikipedia.org/wiki/List_of_localities_in_Wales_by_population
- Four Square (will be used with a free account): https://foursquare.com/
- List of schools in Wales including their 2019 rating: https://gov.wales/sites/default/files/publications/2020-02/national-school-categorisation-system-support-categories-2019-v2.xlsx
- Average house prices on government website: http://publicdata.landregistry.gov.uk/market-trend-data/house-price-index-data/Average-prices-Property-Type-2020-04.csv?utm_medium=GOV.UK&utm_source=datadownload&utm_campaign=average_price_property_price&utm_term=9.30_19_08_20
- Counties in Wales: https://www.townscountiespostcodes.co.uk/counties-in-wales/

Solution to the problem

The cleaned list of localities will be geolocated and categorised using Four Square data and the K-Means model. All good and excellent secondary schools will be geolocated. The classification results and shown on a map will help to narrow down the search to the similar towns and villages only. Finally the average hoes price presentation can be used for the affordability check.