Opening a pub in Aberdeen, Scotland, UK

Bodø, 20.12.2019

Aleksander Jenssen

Picture of the city centre of Aberdeen, Scotland, UK. Source: ANSA.

1.0 Introduction

This report is a part of the final deliverables of the capstone project of the Applied Data Science Capstone course from IBM, delivered through Coursera. Through the following chapters I will present a business problem, describe the data which will be used to find a prospective solution to the problem, describe the research methodology used during the capstone project, present my findings, before finally discussing the findings in light of the business problem and concluding on a possible solution.

1.1 Background

The city of Aberdeen is the third most populated city in Scotland, with a population of 227,560 in 2018, according to the Aberdeen City Council (2019). Being regarded as one of the best cities to start a new business in the United Kingdom, due to it´s high survival rate among newly established enterprises, Aberdeen has become a cradle for economic growth in Scotland.

One category of businesses which are abundant in Aberdeen are pubs. The vast majority of these pubs, however, are located close to the city centre, with few pubs situated at the outskirts of the city. This opens up a business opportunity for prospective entrepreneur who are interested in running their own pub in Aberdeen, if they are able to find a suitable area to start their business. A premise for succeeding in a new venture is the ability to attract a sufficient amount of customers to your business, which means that there are certain criteria that needs to be met when determining whether a given area is suitable for starting a business.

In the case of opening a pub in Aberdeen there are two main criteria that should be met, namely that:

1. *There should be few other pubs in the area, and;*
2. *There should be a sufficient amount of potential customers in the area.*

With these criteria in mind, a business problem has been outlined, as presented in the following sub-chapter.

1.2 Problem/research question, and interested stakeholders

With the aforementioned business opportunity in mind, the research question that this capstone project aim to answer is: *In which area of Aberdeen should an entrepreneur open a pub if they seek to minimize the number of other pubs in the area, while maximizing their reach to customers?*

Finding a solution to the business problem/opportunity, by answering the research question, could be of great interest to several stakeholder. The two most prominent stakeholders in this regard are the potential entrepreneurs who can achieve their goal of successfully starting and running their own pub, in addition to the inhabitants of the selected area(s), who then will have access to pubs in their local area and therefore will not have to travel to the city centre for that particular purpose.

2.0 Data

The purpose of this chapter is to introduce and explain the data that is being used during the capstone project. In the following sub-chapter I will introduce, explain and discuss which data was necessary in order to find a prospective solution to the research question, how and where the data was gathered, how the data was pre-/processed and finally used and analyzed.

2.1 Necessary data

In order to perform the necessary data analysis to determine where one should open a new pub in Aberdeen, in order to minimize the number of nearby pubs, while maximizing the number of potential customers in the area, there are certain types of data which are necessary. First, data containing information about the different areas in Aberdeen is needed. This dataset would ideally contain information about Aberdeen´s neighborhood, their respective populations, and location data in the shape of coordinates. In addition to this data, information about different venues in each area of Aberdeen is needed to properly answer the research question and give recommendations on where to open a pub.

2.2 Data sources

A rather large amount of information about the areas of Aberdeen can be found in csv-format here, at doogal.co.uk, a website which provides public domain licensed data on postcodes in the United Kingdom. This dataset includes area/ward names, coordinates, respective districts and population, among other things. However, the quality of the population data in this dataset was not deemed of sufficient quality to be used in the data analysis. Therefore, this dataset from Aberdeen City Council was used to generate the correct population for each area. Location data from Foursquare API was used to explore venues in the different areas of Aberdeen. Finally, having selected the best area (ward) to start a new pub in Aberdeen, this dataset on Aberdeen´s neighborhoods was used to determine the best neighborhood in the chosen area/ward, according to previously mentioned criteria.

2.3 Data preprocessing

The first major step in preprocessing the different datasets was to load them all into my Jupyter Notebook. Since the population data from Aberdeen City Council was stored in pdf-format, I first had to write it into an excel-file, Population_data.xlsx, containing one row for each area/ward (13 in total) which also displayed the local population, before loading it into the notebook. A similar process was done to load the neighborhood data of Aberdeen towards the end of the capstone project, when all that remained was to determine the best neighborhood to set up a pub, having already determined the best area/ward (see Note/clarification below).

After loading the datasets containing information about Aberdeen´s areas and population into the Jupyter Notebook, they were saved to the variables *ab_data* and *ab_pop*, respectively. Next followed the different stages in preprocessing the *ab_data* dataset: having created the ab_pop dataset myself, there was no need for preprocessing.
First, I examined the shape of the dataset, and discovered that it consisted of 17,064 rows/entries and 46 columns/features. The majority of these features are redundant with respect to the data analysis of this project, so the dataset was reduced to the following features:

1. District
2. Ward
3. Latitude
4. Longitude

To make a single entry for each area/ward I decided to group the dataset by 'Ward' and 'District', and to use the mean value of latitude and longitude as a proxy for locating the geographic center of each area/ward.

Next, I decided to narrow the dataset down to all areas located in the district called Aberdeen City, since the original dataset contain information about postcodes in all of Aberdeen county. The resulting dataset contained 13 rows and 4 columns, a significant reduction. Furthermore, this now meant that my two datasets could be merged, since they both consisted of a single entry for each ward in Aberdeen. The resulting dataframe, *ab_popdata*, now contained one entry for each of Aberdeen city´s 13 wards/area as well as their corresponding districts, latitudes, longitudes and populations.

Note/clarification

Up until this point I have been talking about areas of Aberdeen, occasionally speaking of Aberdeen ´s wards/areas. In Aberdeen a ward represents a voting district for use in elections. I have decided to use Aberdeen´s wards as a proxy for the city´s areas, mostly because of the fact that information regarding Aberdeen´s wards was readily available. I believe that this is justified, in the context of this capstone project, because of the relatively small geographic sizes of these wards: one ward typically is made up of two to five neighborhoods. At the end of the capstone project, when determining the ideal area to open a pub in Aberdeen, I will first determine which ward is the most suitable, before determining the best neighborhood in that ward, based on the criteria mentioned in the introduction.

2.4 Data usage/analysis

The dataset(s) presented above will be used to make a suggestion for the most suitable area(s) in Aberdeen to start a new pub. In order to determine an appropriate location, I will implement machine learning, specifically k-means clustering, to separate areas with an abundance of local pubs from areas with few or no pubs nearby. If one or more areas appear to be equally well suited, other measures will be taken into consideration. In this regard, the criteria highlighted in the previous chapter will act as a guideline. Should there still be uncertainties related to determining which area is preferable, yet new information must be explored. This is a topic for the discussion chapter presented later in the report.