IBM Data Science Capstone Project

Identifying the town in derbyshire most suitable for a new coffee shop

2020

**Introduction**

The rural county of Derbyshire, England is home to the Peak District national park. The Peak district national park was Britain’s first and the world’s second most visited national park [1].

Set on the boarders of the populous cites of Derby, Manchester and Sheffield the peak district is visited by many day trippers wishing to get out and about in the great outdoors as well as investigates Britain’s past with visits to its historic market towns.

The large number of visitors means that the towns situated within Derbyshire are filled with tourist shops and traditional petit cafes and restaurants. However, this has meant that modern new enterprises have often ignored Derbyshire as a place to start a new business. This has left a gap in the market for a new modern, hipster-oriented coffee shop in Derbyshire. It is therefore the aim of this project to identify which town in Derbyshire would be most suited for this modern coffee shop.

The potential locations include Ashbourne, Bakewell, Belper, Buxton, Chesterfield and Glossop. The target market for the coffee shop is 20 to 40-year olds with large disposable incomes.

**Description of Data**

To solve this problem the following data will be required:

* Demographics and population data of the towns to determine the size of the target market in each town.
* Salary data for each town – to determine the amount of disposal income the residents of each town has and therefore determine which towns residents would be most likely to spend the most money in the new coffee shop.
* Location data of existing coffee shops within the towns to determine which areas of the town would have the least competitors.

The data will be sourced from the office of national statistics and Foursquare location data. Further details are highlighted below:

* Salaries and demographics data were sourced from the office of National statistics available at: <https://www.ons.gov.uk/>
* Data downloaded as excel files and includes demographics and salaries of areas of Derbyshire. Unfortunately, data was not available on the individual towns. Therefore, analysis will have to be performed on the areas of Derbyshire and then this data be used for the respective towns.

**Salary Data**

* Excel file containing salary data consists of two sheets. One sheet containing the data and another sheet containing the Meta data. The salary data consists of:
  + Area
    - Amber valley
    - Bolsover
    - Chesterfield
    - Derbyshire Dales
    - Erewash
    - High Peak
    - North East Derbyshire
    - South Derbyshire
  + Geography Code (Unique identifier of each area)
  + Statistics (percentile of salaries, mean and median salary)
  + The data is representative of all genders, all working patterns (e.g. part-time, full-time, etc.) for residents of the areas within Derbyshire.
  + The salary data is denominated in sterling.
  + The data if from 2019.

**Demographics Data**

* Excel file containing demographics data consists of 13 sheets. The sheet MYE2-ALL containing Population Estimates for persons by single year of age and sex for local authorities in the UK, mid 2018 will be used in this analysis. The data consists of:
  + Geography Code (Unique identifier of each area)
  + Area (including the areas of interest which are listed below)
    - Amber valley
    - Bolsover
    - Chesterfield
    - Derbyshire Dales
    - Erewash
    - High Peak
    - North East Derbyshire
    - South Derbyshire
  + Administrative geography areas in the UK.
  + Number of people of each age from 0 to 90.

**Location Data**

* All coffee shops within 1km of the chosen towns centre will be found.
* For each venue the foursquare API will be used to capture the longitude and latitude.

**Methodology**

The method taken for this project is as follows:

1. The salary demographics data was downloaded from the office of national statistics website as two excel files.
2. Data was imported into a data frame from an excel file.
3. Data was visually inspected to understand it contents.
4. Unneeded data was dropped from the data frame.
5. Demographics data was combined into age groups instead of individual years. To make the data more manageable to analyse.
6. Obtain median salary of each region of Derbyshire. Median salary was used instead of the mean salary for the region as the mean value could be heavily distorted by a few very high salaries.
7. Two data sources, demographics data and salary data were combined into one data frame.
8. Regions of Derbyshire was replaced in the data frame by the towns of interest.
9. Obtain coordinates of the towns using the geopy library
10. Cluster the towns into two groups based on the demographics of the town and its median salary using k-means clustering. K-means clustering was used to identify which towns were most suitable for the new coffee shop e.g. towns with young demographics and high salaries.
11. Plot the clusters on to a map for visualisation.
12. Identify which cluster contains the more favourable towns for setting up a new coffee shop.
13. Use the foursquare API to identify other coffee shops within 1km of the town centre for these towns.
14. Plot these coffee shops on to a map.
15. Determine which coffee shop has the least competition.

**Results**

Figure 1: Dataframe Used for k-means clustering displays the demographic data and median salary fields used for k-means clustering. The towns have been separated into to cluster groups.



Figure : Dataframe Used for k-means clustering

Figure 2: Dataframe with Town Coordinates contains the longitude and latitude that is required to plot markers on a map for the town of interest.



Figure : Dataframe with Town Coordinates

Figure 3: Map of Clustered Towns displays the two cluster groups for the 6 towns that were analysed. Cluster 0 containing Ashbourne, Bakewell, Buxton and Glossop are represented by a red marker and group 1 containing Belper and Chesterfield are represented by a purple marker.

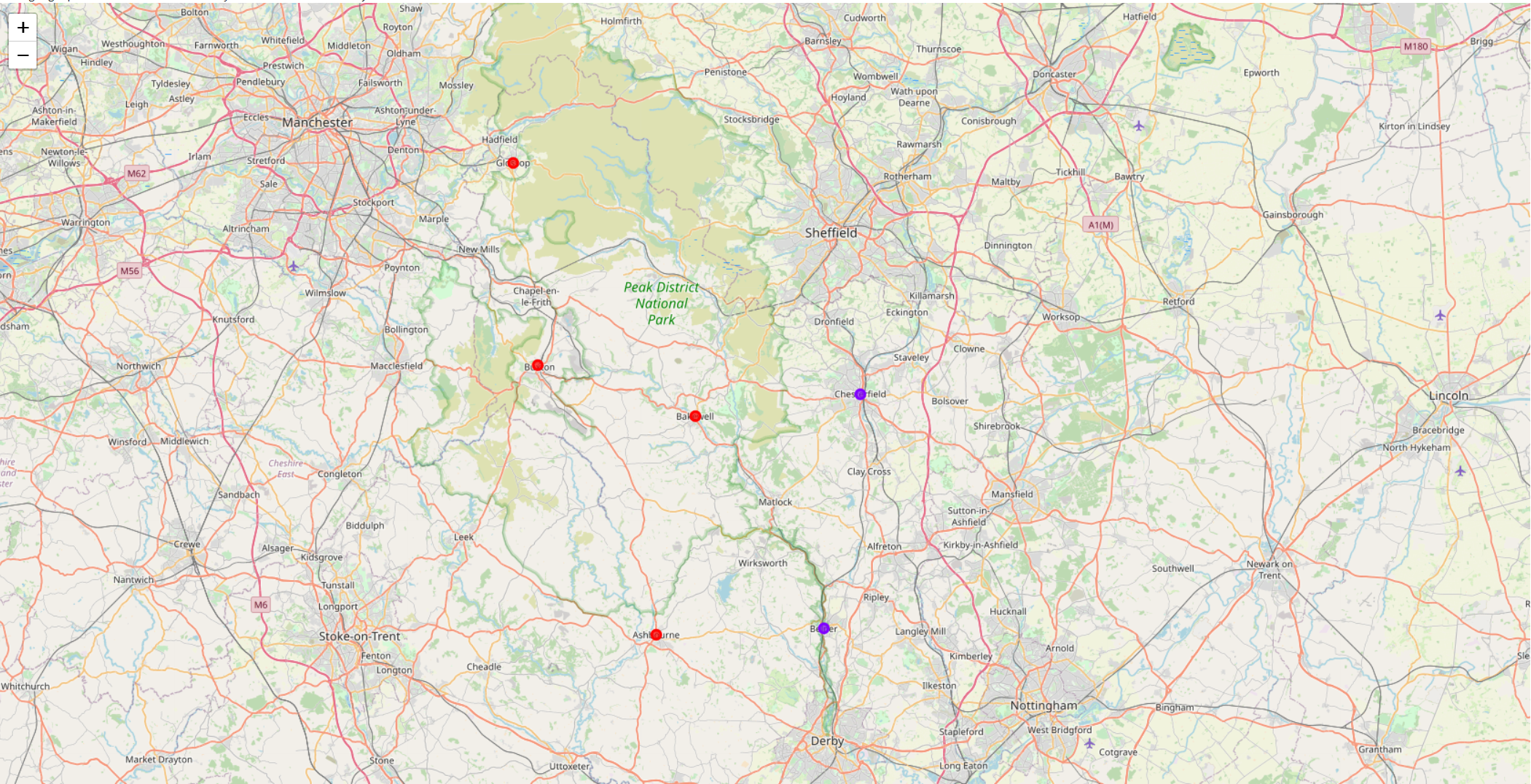


Figure : Map of Clustered Towns

Figure 4: Map of Coffee Shops in Chesterfieldshows the location of coffee shops within 1km of Chesterfields town centre.

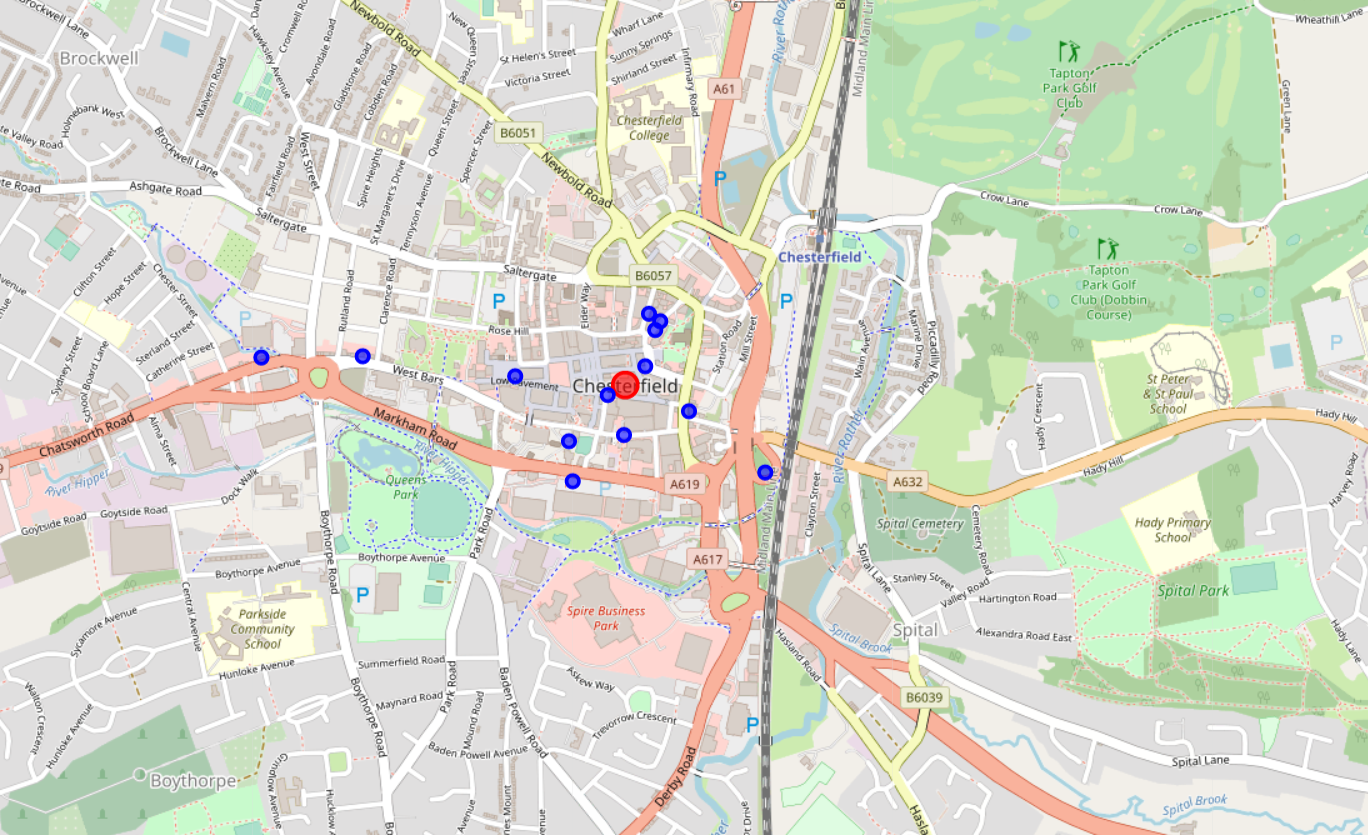


Figure : Map of Coffee Shops in Chesterfield

Figure 5: Map of Coffee Shops in Belpershows the location of coffee shops within 1km of Belper’s town centre.

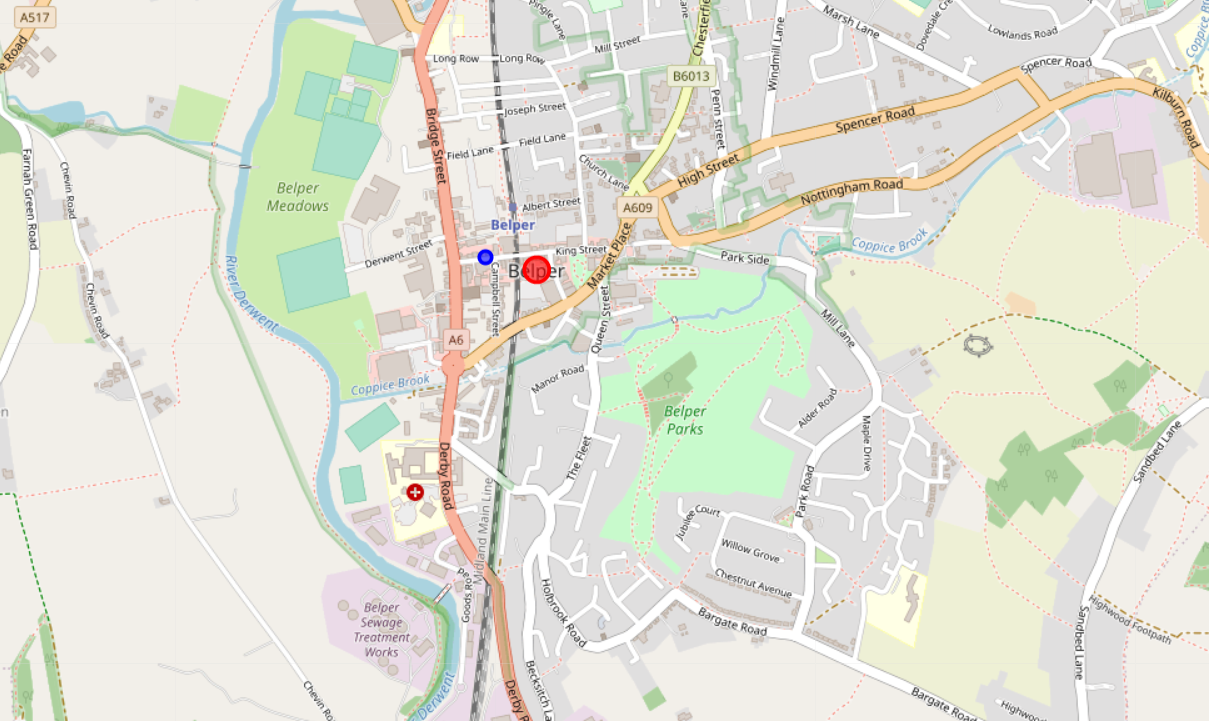


Figure : Map of Coffee Shops in Belper

**Discussion**

The results from this data investigation have shown that Belper appears to be the most suitable town for setting up a new hipster coffee shop.

Belper’s young demographics and the small number of existing coffee shops has proven to be the most influential factors in Belper being the most optimal town for the new coffee shop.

The demographics between the 6 towns varies drastically. However, this variation is greater in the younger age categories. Belper and Chesterfield have approximately double the number of people in the age category 16-25 whereas Belper and Chesterfield have approximately 25% more people in the older age categories. This shows that the demographics of Belper and Chesterfield is younger than the other towns. In addition, the wage difference between the towns is marginal and therefore is less important. However, it should be noted that this data is for regions within Derbyshire and not the individual towns therefore this data is not truly accurate but is the only demographics and salary data available for these towns.

The younger demographics of Belper and Chesterfield and the similar salaries between all six towns meant that the k-means clustering algorithm identified that Belper and Chesterfield should be in the same cluster.

After the segmentation further investigation was performed to determine how many coffee shops already exist in the two towns. Chesterfield already has 12 coffee shops within 1km of the city centre. Belper only has 1 coffee shop within 1km of the town centre. Therefore, it stands that new coffee shop in Belper would face less competition and as such Belper was identified as the town most suitable to start a new coffee shop.

Further investigation work should be performed to decide if Belper is suitable for a new coffee shop or not. This would include identifying if there are shops other than coffee shops such as tea rooms which may provide competition to the new shop. Also, the location within Belper for the new coffee shop has not been identified. To identify the area of Belper most suitable for the new shop data such as the number of people who walk past on the street and risk of flooding or other natural disasters should be investigated. Also, it is still important that the town be inspected in person to identify any opportunities or challenges that may be present that the data has not identified.

**Conclusion**

In this piece of work Belper has been identified as the most suitable town in Derbyshire to set up a new coffee shop. Demographic and salary data have been used to determine a towns suitability for the new coffee shop. K-means clustering was used to segment the towns into two clusters. The towns in the cluster that displayed the characteristics that were initially identified as favourable were further investigated to determine the amount of competition in these towns. The town with the least completion was chosen as the town to set up the coffee shop. This town was Belper.

Further work should be conducted before confirming to start a coffee shop at Belper this should include but not be limited to: identify the area of Belper most suitable for the new shop data such as the number of people who walk past on the street and risk of flooding or other natural disasters should be investigated. Also, it is still important that the town be inspected in person to identify any opportunities or challenges that may be present that the data has not identified.

**References**

[1] - <https://www.holidaycottages.co.uk/blog/top-10-things-to-do-in-the-peak-district>