|  |  |
| --- | --- |
| NEW CINEMA LOCATION | |
|  | Film Reel Stock Photos, Pictures & Royalty-Free Images ... |
| BATTLE OF THE NEIGHBOURHOODS CAPSTONE DATA SCIENCE PROJECTAugust 2021Catherine Brown |  |

|  |
| --- |
| INTRODUCTION *Scenario:*  I have been tasked by a client with finding a suitable location in the Nottinghamshire area for a new cinema to be built or for an existing building to be converted. The Nottinghamshire area already has many cinema venues, so the location has to be carefully selected to ensure that it will be a successful business venture. My client already has a chain of independent cinemas in the North of England and is looking for a new location in the Midlands, with Nottinghamshire being the preferred location. Other individuals or businesses interested in opening a new Cinema in the Nottinghamshire area may also be find this project of use.  *The scope of this project within the wider business analysis:*  Opening a new cinema requires significant capital investment. Deciding on a location, whilst an important part of the strategic decision, is only one of many factors to be explored. The local demographics of the area is an important factor to consider. Other analysis such as SWOT and 5 Market forces should also be a starting point for decision making. The barriers to entry for building a new cinema are high due to the upfront capital investment and the understanding of the business sector. Streaming providers such as Netflix and Amazon Prime should be considered as a threat to the viability of existing and new entrants to the cinema market. The high supply cost of showing new release Hollywood movies should be considered when deciding what type of cinema to build. The projected revenues and costs should then be computed to determine the actual return on investment to ascertain whether the project is potentially viable. Funding would then need to be sourced. The scope of this project is to determine only which NG (Nottinghamshire) postcode is the most suitable location for a new cinema to be built.  *Types of cinemas:*  The existing cinemas in Nottinghamshire are Multiplex cinemas such as Odeon and CineWorld, along with smaller chains such as the Savoy cinemas. There is also a popular independent cinema, Broadway, in the centre of Nottinghamshire that specialises in classics and alternative films. The type of cinema to be built would depend on the results of the business analysis. Cinemas such as the Savoy or Broadway are much smaller than the Multiplex cinemas and will perhaps show only a couple of the main release films at one time, offering older films that are cheaper from a supply perspective on the other screens. They offer a more intimate experience than the multiplex cinemas, with Broadway also having a bar and small restaurant. Creating more of an experience is likely to be ever more important when competing for customers with Netflix and Amazon Prime. DATA USED *1. NG Postal code data:*  The Nottingham postal code data will be extracted from the webpage: <https://en.wikipedia.org/wiki/NG_postcode_area>’, using the function dfs=pd.read\_html(url) to read the data into a pandas dataframe (df). This data will be used to select from when determining a suitable location for a new cinema. The extracted data will need to have coordinates found, so the postcode district and post town are combined into a full address column:     1. *Converting the NG Postal code data into coordinates for mapping:*   Using Geopy, the NG Postal code data will be used to find the approximate centre latitude and longitude coordinates of the postal codes. These coordinates will be added to the dataframe.     1. *Mapping the NG Postal codes*   The postal codes will then be mapped using Folium to provide a visual of the distribution of the postal codes.   1. *Obtaining existing Cinema venues around the Nottinghamshire area*   Foursquare will be used to obtain existing cinema venues around the Nottinghamshire area. This information will be stored in a dataframe, the data cleaned and the resulting location venues will be mapped onto the same map as the NG Postal codes. This will provide an initial visual of postal areas that do not already have a nearby cinema venue. It is better to identify a location that does not already have a cinema so that there is less competition for customers.   1. *Clustering the NG Postal codes using popular venues*   Foursquare will be used again to find the nearby 100 venues for each postal area. The data is then grouped by postal code area and sorted into the top 10 venues within each postal code area. A clustering method called KMeans will be applied to this data in order to cluster the postal code areas based on top local venues. To find the optimal k for the algorithm, the Elbow method is used. The information produced from this clustering exercise will provide an idea as to which postal code areas are similar types of Neighbourhoods based on nearby venues.   1. *Mapping the clusters on the same map and determining the best location*     With each area assigned to a particular cluster and imposed on the map, this will then provide an additional visual of which clusters tend to have cinemas and help identify if there are any postal codes that fall in the same cluster but that do not have a nearby cinema. Such a location could be a potentially good location for a new cinema.   1. *Limitations of the scope of this project*   It is important to note again that proximity to existing cinemas and similarity of neighbourhoods based on popular venues are only a very small part of the overall picture when deciding where to build a new cinema in Nottinghamshire and that the results of this project should be used in conjunction with a vast array of other supporting material as discussed in the wider business analysis section.  **METHODOLOGY**  The following methodology was followed in this project:   1. The Nottingham postal code data was extracted from the webpage: <https://en.wikipedia.org/wiki/NG_postcode_area>’, using the function dfs=pd.read\_html(url) to read the data into a pandas dataframe. In order to use this data to find the coordinates of the postcode area, it was necessary to combine the potcode district and post town into a full address. This full address column was then added to the dataframe:      1. Geopy was used to find the latitude and longitude coordinates of the postcode districts. Two of the postcode districts did not return any information, but Geopy was used to find the latitude and longitude values for these areas (on re-running the code, Geopy did not work for this), and the NaN data was replaced with the correct coordinates:      1. The postal code data was mapped using Folium and pop up markers:      1. Foursquare was used to find existing cinema venues within a 50,000m radius of the central coordinates of the Nottinghamshire location on Geopy:      1. The results were read in a json file and stored in a dataframe calles df\_fil. The data required cleaning – certain venues found using the search query cinema were not actually cinema venues when the category was checked, but some were coffee shops or bus stops. These were removed from the dataframe to leave only real cinema venues:      1. The existing cinemas were then mapped onto the existing map shown in Step 3, with the large red circle marking the centre of the Nottinghamshire area:      1. An exploration of the different Nottinghamshire postal codes was then made using Foursquare. The following function was created to extract nearby venues:     This function was applied to search the original postal code dataframe from stages 1&2 (df) and a new dataframe was formed (Notts\_venues). The Notts\_venues dataframe was then grouped by Postal code and there were found to be 150 unique categories of venue.   1. One hot coding was used to add the unique categories of venue by instance and postal code to a new dataframe (Notts\_onehot). This dataframe was then grouped by Postal code and the mean frequency of each nearby venue was found (Notts\_grouped). 2. The Notts\_grouped dataframe was sorted to show the top 10 venues by postal code (Notts\_venues\_sorted):      1. A Notts\_clustering dataframe was created based on the Notts\_grouped data above. The unsupervised machine learning algorithm K-Means was applied to the data to group the postal codes into clusters according to their nearby venues:     K-Means forms the clusters by minimizing the sum of the distance of points from their respective cluster centroids.   1. The Elbow method of finding the most appropriate value for k was used:      1. A dataframe Notts\_merged was created to add the cluster labels to the Notts\_venues\_sorted dataframe:      1. The following code was added in order to change the Cluster labels to integers from floats so that they could be plotted on the map:      1. The postal codes were clustered into 4 clusters and the cluster results plotted onto the map using Folium to get a good overall visual image of the best location(s) to build a new cinema.   **RESULTS**  The Postal codes were clustered into 4 clusters, this being the most appropriate value of k for the kmeans clustering, with the top 5 most common venues defining their name and shown as follows:  Red cluster – Sports shop area    Purple cluster – Pubs and Grocery stores areas    Turquoise clusters – Pubs, bars and restaurant areas    Green cluster – Brewery and Ice cream shop area    The results of the clusters being mapped is as follows:  **DISCUSSION**  The map above shows: the Pop up markers containing the Postal codes for the NG postcodes with the corresponding cluster area (out of a possible 4 clusters) shown by the coloured circle at the point of the marker. The map also shows the existing cinema locations marked in dark blue and the centre of the Nottinghamshire area is marked with the large red circle.  This analysis has done the following:   * Taken the Nottingham postal code data and found cinema locations within a 50,000m radius of the Nottinghamshire county central latitude and longitude coordinates. * Both the postal codes and the nearby existing cinema venues have been plotted on a map of the region. * The postcode areas have been explored to examine nearby venues and the top 10 venues in a 4,000m radius from the postcode coordinates have been found. * The postcode areas have been compared and clustered using the kmeans algorithm method. * The resulting clusters have been plotted as an overlay onto the cinema and postcode map. * The resulting visual provides a good indication of what type of clusters already have a cinema and therefore which areas may be a good location for a new cinema to be built.   The results above are limited in that they have used only nearby cinemas and clustering of similar Neighbourhoods to determine a suitable location. As can be seen from the results of the clustering, the Neighbourhood clusters in Nottinghamshire are fairly similar - they all contain a lot of pubs, although I have tried to name them according to their most popular venues. There is good coverage with regards to existing cinema venues, with the largest concentration of venues being in the centre of Nottingham as one may expect. Grantham (NG31 and NG32), Long Eaton (NG10) and Newark (NG24) are all found in the purple clusters which has pubs and grocery shops as being the most popular venues. Sleaford (NG34) falls into the purple cluster but does not have a nearby cinema.  **CONCLUSION**  I recommend that the area of Sleaford is explored further as a potential location for a new cinema to be built or for an existing building to be converted since it has similar other venues in the area to other towns that already have cinemas. It also does not have a cinema in a nearby location.  Since the map shows that Sleaford is not a major city, it is likely that the size of the cinema should be relatively small and further revenue may be gained from providing a food and beverage option, since the most popular venues in the area are a supermarket, a fast-food restaurant and a gastropub. |

**Week 2 Task (to be completed)**

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