# Determining the ideal development sites within Cambridge, UK

Christopher Winnard

30 April 2020

## Introduction

### Background

The city of Cambridge, UK is a rapidly expanding City. With direct and easy commuter links to London many are looking to this historically significant and scienic area. On top of this, the industry sector within the city itself has been expanding. The volume of jobs for young working professionals is increasing, this is increasing the volume of this demographic moving into the city.

Property developers and building companies are looking to provide new properties aimed at this demographic and so are looking for the ideal areas to develop within. To gain the most interest in their properties, as well as being able to charge more for more desirable properties, the developers must select candidate areas that fit certain criteria. The ideal sites would be close to the train stations to allow for commuters. Have lower reported crimes and crime rates, allowing the developers to claim they are built within ‘Safer’ areas. Given the target demographic, the developers are also looking to build within areas that contain venues suited to the investors, these include restaurants, bars/pubs and cafes/coffee shops, as these are desirable venues for an area.

The problem is that data that could help the developers decide which areas they should invest in is located and stores within many different sites, for example in government records, local police datastores and other online databases.

## Data Acquisition

The data to be analysed within this project is stored in many locations. The post code location data can be found here: <https://www.freemaptools.com/download-uk-postcode-lat-lng.htm>. This webpage contains a csv file documenting all the registered postcodes within the UK. This encompasses over 190,0000 locations, each of which contains longitude and latitude co-ordinates.

Crime statistics must be provided by each individual constabulary (this is the term for the regional police forces) on a monthly basis. Data for each month can be downloaded from the government portal found here : <https://data.police.uk/data/> . Files containing the data for the last 6 months was collected.

Foursquare Developers API was utilised to collect information on venues within a chosen demographic area based on specified search terms and co-ordinates.

## Data Formatting and Cleaning

UK postcode dataset was downloaded and a dataframe created with the input. Post codes are within the format Letter, Letter, Number, space, Number, Letter, Letter, for example CB1 1AA. This dataset was filtered to remove all postcodes that didn’t relate to the target area, Cambridge, denoted with the area code CB. This file still contained over 10,000 individual locations, each with specific geographic co-ordinates. These were clustered into 15 sub-groups: CB1, CB2, CB3, CB4, CB5, CB6, CB7, CB8, CB9, CB10, CB11, CB21, CB22, CB23, CB24, CB25 and the locations of all were used to determine a mid-point for each ‘Area’. The 5 locations closest to a central location for the city were selected for further analysis and a new dataframe containing only co-ordinates for these 5 Areas was created. The co-ordinates for the 2 train stations within the city were also determined using the foursquare API for analysis in relation to the individual areas.

The crime statistic files were downloaded for the last 6 months. These 6 files were merged and a datframe created containing information on all crimes reported within the country of Cambridgeshire. The file contained information on the exact geographical location (Longitude and Latitude data), type of crime, street name/location, data of incident and region within county. This file was filtered and a new dataframe created containing only the crimes within the City of Cambridge’s boundaries. Redundant columns were removed resulting in a data frame containing only the co-ordinates of the crime, type of crime and the police specific location denotation. Any entries that contained missing geographical data were removed from the dataset.

Foursquare API was used to request information on the top 100 venues and top 10 for each of our assigned Areas. The resulting information was converted to a pandas dataframe. One hot encoding and k-means analysis allowed me to cluster the venues locations. The venue categories were then analysed and subgrouped created to encompass multiple venue types. These included:

Restaurants: American Restaurant, Asian Restaurant, Buffet, Burger Joint, Chinese Restaurant, English Restaurant, Fish & Chips Shop, Gastropub, Indian Restaurant, Italian Restaurant, Pizza Place, Portuguese Restaurant, Restaurant, Steakhouse and Thai Restaurant

Bars/Pubs: Bar, Beer Garden, Brewery and Pub

Cafe: Bakery, Breakfast Spot, Café, Coffee Shop, Deli / Bodega, Ice Cream Shop, Lounge and Sandwich Place

Recreation: Botanical Garden, Field, Gym, Gym / Fitness Center, Gym Pool, Park, Scenic Lookout and Soccer Stadium

Arts: Art Gallery, History Museum, Indie Movie Theater, Performing Arts Venue, Science Museum and Theater

Shops: Bookstore, Clothing Store, Food & Drink Shop, Furniture / Home Store, Garden Center, Grocery Store, Market, Pharmacy, Record Shop, Supermarket and Wine Shop

Other: Bed & Breakfast, Hotel, Multiplex, Office and Train Station

The resulting categories were analysed for each of the areas and plotted for visual analysis.

## Methodology

Foursquare API was used to determine center point co-ordinates for our selected city and well as for train stations located within the city. Uk Postcode locations were collected and clustered and a centre point location for selected regions created, termed ‘Areas’, designating areas within our selected city. Folium library was imported and interactive maps created with markers denoting our cities ‘Areas’. The Geopy function ‘distance’ was utilised to calculate the exact distance between our clustered areas and the 2 train stations located within the city. This information was plotted using the Matplotlib .plot function for visual representation.

Crime statistics data was collated, clustered and filtered prior to mapping onto our selected city using Folium maps. A heatmap was created to visually interpret the crime levels according to our selected ‘Areas’.

Foursquare developer’s API was utilised to collect information on the top 100 venues within our chosen city. The resulting information was converted to a workable dataframe and locations of each venue visualised on an interactive folium map. The resulting dataframe was analysed using the sklearn library function kmeans, one hot encoded and k-means were applied to the dataset and the resulting clusters were plotted onto a folium map for visual analysis of where the clusters lied. Venue categories were clustered into subgroups, as stated in the previous section, and one hot encoded. The resulting dataframe was visually analysed using the Matplotlib .plot function as bar-charts displaying the number of venues per ‘Area’ to discern which areas contained the most of which types of venues.

## Review and Conclusions

Firstly we gathered geospacial data on all the postcodes within the UK, provided by the royal mail postal service. This data was restructured and used to determine central points for candidate districts within our chosen city. These were filtered down to 5 potential areas and visulised on an interact map. The distance of these locations from the Cities 2 train stations was determined and visualised, both graphically and on an interactive map.

We then gathered information on reported crimes within the city. The local police force provides crime statistics monthly for each county, this also contains location data of the reported crimes. Data was collected from the past 6 months and merged into one workable dataframe. This was further processed to provide information just within our selected areas. This crime data was visualised according to location and an interactive heatmap was created to demonstrate which areas have the highest concentration of crimes reported.

We then collected information on the local venues within Cambridge using Foursquare API. The different establishments were clustered using one hot encoding and K-means analysis and mapped to provide visuals for our different clusters. The Venues types were sorted and sub-grouped into defined subsets which allowed us to analyse and compare the different types of venues within each area. This information was displayed graphically to allow the investors to determine which types of venue may be most desirable to their targeted demographic allowing them to make a more informed choice on which locations may be more appropriate to develop on.

Based on the crime statistics data the 'safest' location to develop properties would be within either the CB3 or CB2 areas. CB3 is far away from both of the train stations within Cambridge whereas CB2 is very close to the Central station (but far away from Cambridge North train station). When looking at the top 100 venues for each area, CB2 has the highest number of Restaurants and a very high proportion of Cafes. When looking at the closest 10 venues CB2 doesn’t have the most of any one venue type, however it does have a good balance of all venue types, except for shops.

The data produced from this project suggests CB2 would be one of the best potential development sites for our chosen demographic based on the desirable traits listed at the beginning of the project. However futher data should be analysed to confirm this hypothesis. This could include breaking down the crime statistics into crime type eg Violent and mild. Housing prices for each Area should also be analysed and this is a very important factor to clients when selecting a property to purchase and will also impact the retail price of any properties build by the developers.