Identification of Ideal Locations for Student Accommodation

Nikesh Lama

August 2021

1 Introduction

1.1 Background

Nottingham is a city in central England's Midlands region in the United Kingdom. There are around 65,000 students at Nottingham's two universities - University of Nottingham and Nottingham Trent University. With a huge student population, Nottingham is one of the most vibrant cities in the UK. Nottingham is ranked as the 6th best city in the UK for students and 48th in the world, according to the QS Best Student Cities 2019. Due to a large influx of students at the beginning of each academic year, university halls of residence are not enough to accommodate all the students. Also, most students prefer to seek residence via private student accommodation due to affordable rent and wider options in terms of location and house mates. This opens up business opportunity to invest and explore private accommodation services to the university students ensuring maximum safety, accessibility to amenities and minimising distance to the University.

1.2 Business Problem

***Business Problem**: Identify ideal locations around Nottingham Trent University (NTU) such that the locations have access to wide range of amenities, are safe and are at close proximity to the university (within 5 km).*

In this project, I utilise Foursquare API to explore neighbourhoods around NTU to provide consultation for the best locations for investing in student accommodation. The solution provided will be useful for business owners to choose locations around NTU to provide accommodation services to the university students. Mainly, availability of facilities around the locations and the number of criminal events reported are taken into account. Usually, students prefer to live in a close proximity to university campuses so the locations are restricted within 5 km radius of NTU.

2 Data acquisition and processing

2.1 Datasets

The two main datasets that are used are geolocation postcode data for Nottinghamshire and the latest corresponding crime data from June 2021.

- **Postcode data for Nottingham** : Geolocation dataset for NG postcodes in Nottinghamshire was downloaded from NG postcodes . The dataset was further cleaned and narrowed down with only necessary fields before using with Foursquare API.
- **Nottingham Crime Data** : Crime data is a publicly available data downloaded from data.police.uk. To keep the problem simple enough, I computed total number of all crimes reported for each location. This information was then matched for each location from the postcode data.

- **Nottingham Trent University Data**: Nottingham Trent University postcode is available from the university website . Geocoder, which is a python based geocoding library, was then used to extract coordinates.
- Neighbourhood outcodes i.e. first part of the postcode within 5 km of the university was extracted using https://api.postcodes.io
- Venues around a particular latitude within 500 meters of are explored using Foursquare API. Venue co-ordinates, venue name and venue category were used for further analysis.

2.2 Data wrangling and feature selection

The data often requires cleaning and processing further based on the problem one is trying to tackle. Hence, wrangling the data and preparing a final set of data are essential for data science problems. This section covers data cleaning, processing, organising, modifying and preparing it for analysis.

2.2.1 Postal code data

Postcode data for Nottingham resulted in a table with 49 features and 37,461 entries. However, there were many missing data and we do not need all the features to make suggestions for the best locations. The features that are of interest are [**'Postcode', 'In Use?', 'Latitude', 'Longitude', 'District', 'Postcode district','Ward', 'LSOA Code'**]. 'In Use?' feature shows if the postcode is currently in use or not. The entries were discarded if this field was 'No' and this feature was dropped as all the entries are the same after discarding 'No' entries. Lower Layer Super Output Areas (LSOA) are a geographic hierarchy designed to improve the reporting of small area statistics in England and Wales. LSOA code was used to find areas from the crime data.

The postcodes are divided into two parts: (i) outcode - NGxx (ii) Incode - XXX. Postcodes that are very similar are at close proximity to each other and doesn't bring enough variation in the data. Hence, similar postcodes were categorised as area code and a new feature under 'Area_code' was added to the dataset. The area code is composed of Outcode and first part of the incode. For example: NG1 5, NG11 8, etc. The data cleaning was completed for postcode data which were used later with crime data. The final clean dataframe screenshot is shown in the Figure 1

	Postcode	Latitude	Longitude	District	Postcode district	Ward	LSOA Code	Area_code
0	NG1 1AA	52.955053	-1.141030	Nottingham	NG1	St. Ann's	E01033405	NG1 1
1	NG1 2AA	52.954794	-1.150991	Nottingham	NG1	Castle	E01033406	NG1 2
2	NG1 3AA	52.954591	-1.142989	Nottingham	NG1	Castle	E01033407	NG1 3
3	NG1 4AA	52.955386	-1.149835	Nottingham	NG1	St. Ann's	E01033409	NG1 4
4	NG1 5AA	52.954523	-1.156208	Nottingham	NG1	Castle	E01032522	NG1 5

Figure 1: Cleaned postcode data

2.2.2 Crime data

To measure safety of the area crime data based on the reported crimes from each area is used. To simplify the process, latest data available at this time(August 2021) is for June 2021. The dataframe consisted of 12 features and 12592 entries. The features were [**'Crime ID', 'Month', 'Reported by', 'Falls within', 'Longitude','Latitude', 'Location', 'LSOA code', 'LSOA name', 'Crime type','Last outcome category', 'Context'**]. There were no postcode information but LSOA code can map it back

	Postcode	Area_lat	Area_long	District	Outcode	Ward	LSOA Code	Area_code	Total Crimes
0	NG1 1AA	52.953105	-1.141712	Nottingham	NG1	St. Ann's	E01033405	NG1 1	46
2	NG1 2AA	52.953374	-1.148357	Nottingham	NG1	Castle	E01033406	NG1 2	264
3	NG1 3AA	52.956909	-1.146749	Nottingham	NG1	Castle	E01033407	NG1 3	24
4	NG1 4AA	52.960466	-1.152731	Nottingham	NG1	St. Ann's	E01033409	NG1 4	152
5	NG1 5AA	52.955309	-1.157221	Nottingham	NG1	Castle	E01032522	NG1 5	40

Figure 2: Cleaned crime data

to the postal area. 'LSOA code' and 'Crime type' were extracted to compute total number of crimes reported for each LSOA code.

The next stage was to combine this information with the postal code such that the final postal code data also include total crimes reported for each postal code area. Crime dataframe was merged with postal code data based on LSOA code on both data frames. The final dataframe screenshot is shown in Figure 2.

The total crimes feature was normalized using the *min-max* method. The normalized values were then converted into categorical scores [1: 8] based on the values. I found simply categorising as 'Low','Medium' and 'High' made it difficult to find optimum number of clusters with elbow method. So, I have divided into 8 different scores by diving the normalized values into ranges:

- Score 1: Values < 0.05
- Score 2: Values $>= 0.05$ and Values< 0.1
- Score 3: Values $>= 0.1$ and Values< 0.2
- Score 4: Values $>= 0.2$ and Values< 0.3
- Score 5: Values $>= 0.3$ and Values< 0.4
- Score 6: Values $>= 0.4$ and Values< 0.5
- Score 7: Values $>= 0.5$ and Values< 0.6
- Score 8: Values $>= 0.6$

Most values fall below 0.6 so values above 0.6 can all be categorised into one score. of 8 which is the highest scoring indicating very high reported crimes.

2.2.3 University data

Since I have decided to work with Nottingham Trent University due to its central location, the postcode was extracted from the university website and the co-ordinates were extracted using geocoder API. Based on the latitude and longitude of the university, all the area outcode within 5 km radius were identified and appended into the university data frame. Finally this data frame was merged with the combined dataframe of the postcodes and the crime data. The final merged data frame included all the necessary information to perform analysis on the neighbourhood. The screenshot of the combined data frame is shown in the Figure 3. Postcode_x corresponds to the university postcode and Postcode_y corresponds to the area postcode which are within 5 km radius of the university. Corresponding information for each post code are all the columns after Postcode_y.

	Institute	Postcode_x	Ins_Latitude	Ins_Longitude	Outcode	Postcode_y	Area_lat	Area_long	District	Ward	LSOA Code	Area_code	Crime Score
0	Nottingham Trent University	NG1 4BU	52.958137	-1.154233	NG1	NG1 3AA	52.956909	-1.146749	Nottingham	Castle	E01033407	NG1 3	3
1	Nottingham Trent University	NG1 4BU	52.958137	-1.154233	NG1	NG1 1AA	52.953105	-1.141712	Nottingham	St. Ann's	E01033405	NG1 1	4

Figure 3: Merged data frame with the university data

3 Exploratory Data Analysis

So far, I have discussed how the data are cleaned, processed and new features are created. In this section, I explore the dataset with exploratory analysis before proceeding with the cluster analysis.

Analysing the descriptive statistics of the number of crimes reported shows that the mean crime reported was 28.17 with a standard deviation of 35.81, the minimum number of crimes reported was 1 and the maximum was 264. The distribution is shown with a box plot in the Figure 4. The boxplot distrbution shows that the maximum values are outliers in the plot. These values have significant shift from the distribution indicating disproportionate increase in the number of crimes reported. These areas are not a good choice due to safety concerns.

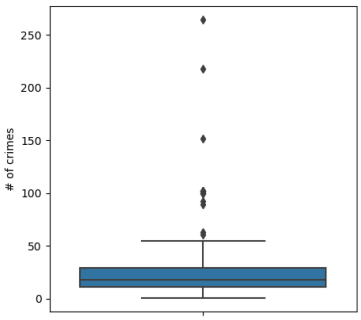

Figure 4: Distribution of the number of crimes reported for each LSOA code

Further, I explored the top 20 areas which have the highest and the lowestr number of reported crimes in Nottinghamshire. Figure 5 shows a horizontal bar chart for top 20 safest and top 20 most crime ridden area codes in Nottingham. As we can see that the top 2 most crime ridden locations have more than twice as much crimes reported when compared to the 4th most crime ridden place. The rest of the area codes have gradual decrement so these areas have comparable criminal activities. Keeping the area safety in mind, the top two locations were discarded which have more than 200 crimes reported in a month.

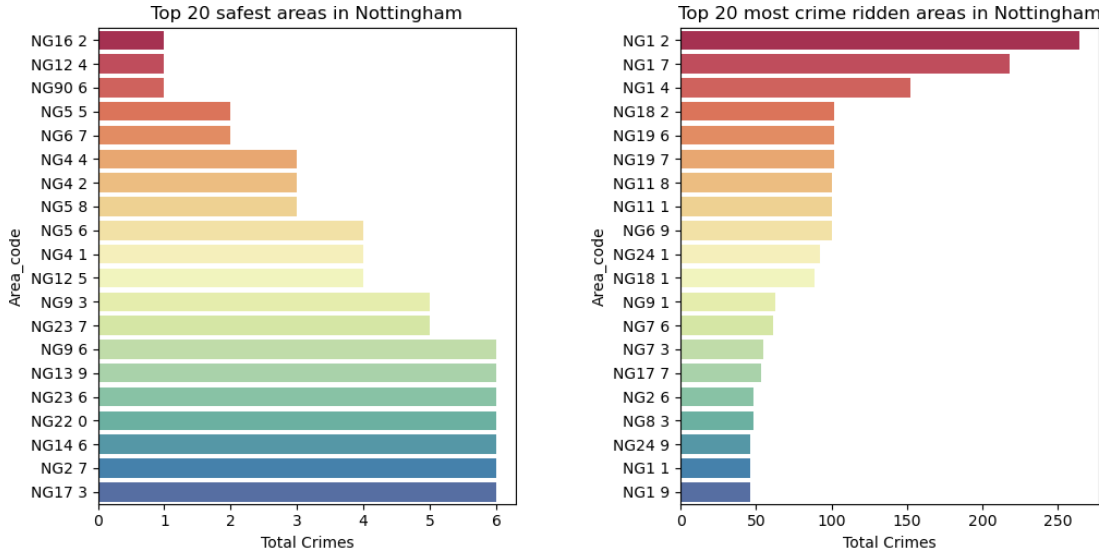

Figure 5: 20 most and least crime ridden areas in Nottingham

As already discussed in the previous section, the reported crime numbers were normalized and then categorised into score based categories[1:8] where 1 represents safest and 8 represents highest criminal activities. Exploring the number of locations that fall into each category is shown in the Figure 6. The figure shows that most regions in Nottinghamshire fall into safer categories (scores 1-5). Most areas fall on the bottom half of the scores ,ranging from 1-4, which shows that Nottingham is mostly a safe place.

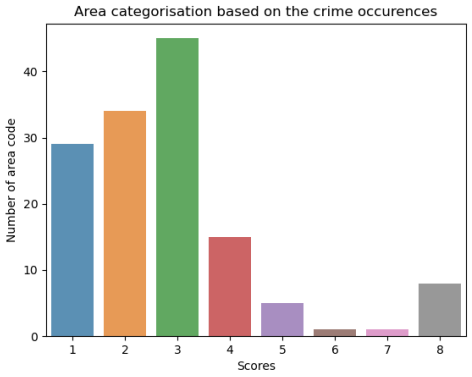

Figure 6: Area categorisation based on the number of reported crimes

I used Folium library for visualizing geospatial data. i.e. location of areas on the map The visualisation helps to (i) validate the locations, (ii) area coverage and (iii) where the area are in respect to the university. Figure 7 shows a geospatial image of Nottingham Trent University and surrounding areas. The left figure shows locations around NTU where each blue circle indicates an area code. The right figure whereas shows a heatmap based on number of crimes reported. Based on the heatmap, there are small pockets of areas with a very high number of crimes reported. The figure helps to obtain a very good idea about how the settlement areas are spread out where the most crimes were reported.

Figure 7: Geospatial map with NTU at the center of the map. A red circle with a radius of 5 km is drawn to show the area of interest around NTU. (Left) Blue markers indicating each area code. (Right) A heat map based on number of crimes reported on each location in Nottingham

After deducting the two most crime ridden areas and narrowing down the locations within 5 km of NTU, 59 area codes were identified. Geospatial map of the identified areas is visualised in Figure 8. The rest of the further analysis were then conducted with these locations to find out which locations would promise the best options to the business owner.

4 Results

Until this point, I have completed data wrangling and used exploratory data analysis to further narrow down the locations for cluster analysis. In this section, I use Foursquare API to explore venues around each location and make further recommendation on which locations are the best for opening a student

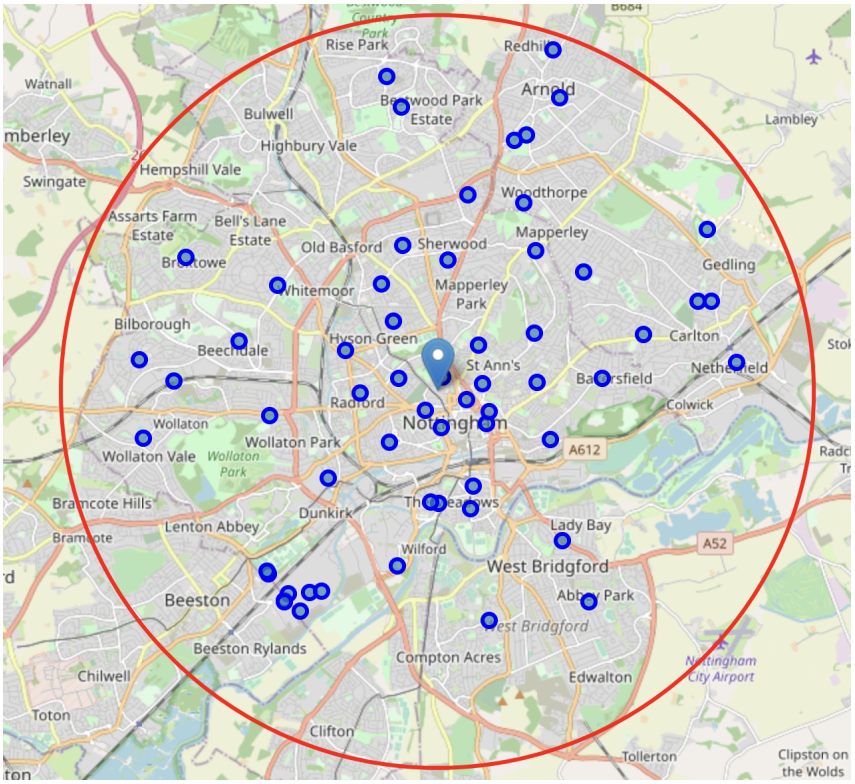

Figure 8: Candidate area codes within 5 km of NTU

accommodation. Availability of venues plays a key factor in student's life so the more venues a location has the better choice that location is for students. Foursquare API returned 152 unique categories of venues. The top 20 venues are shown in Figure 9. The top venues shows a wide array of venues. To perform cluster analysis based on the crime scores and venues, these features were encoded with 'One-Hot' encoding. One-hot encoding converts each type of venue as an individual feature.

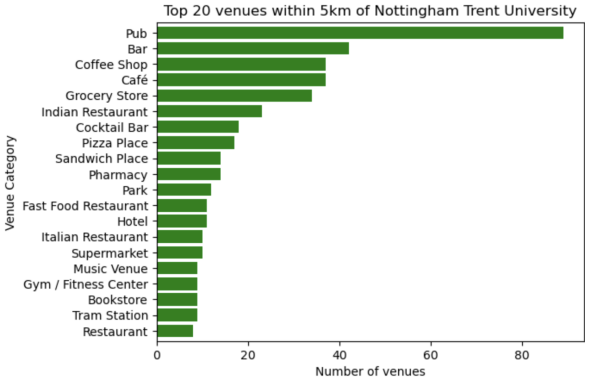

Figure 9: Top 20 venues around NTU

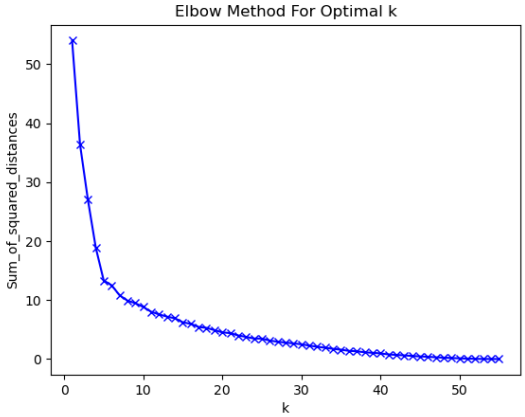

Figure 10: Elbow method to find optimal number of clusters

KMeans clustering was employed to cluster the locations based on venues categories and crime scores

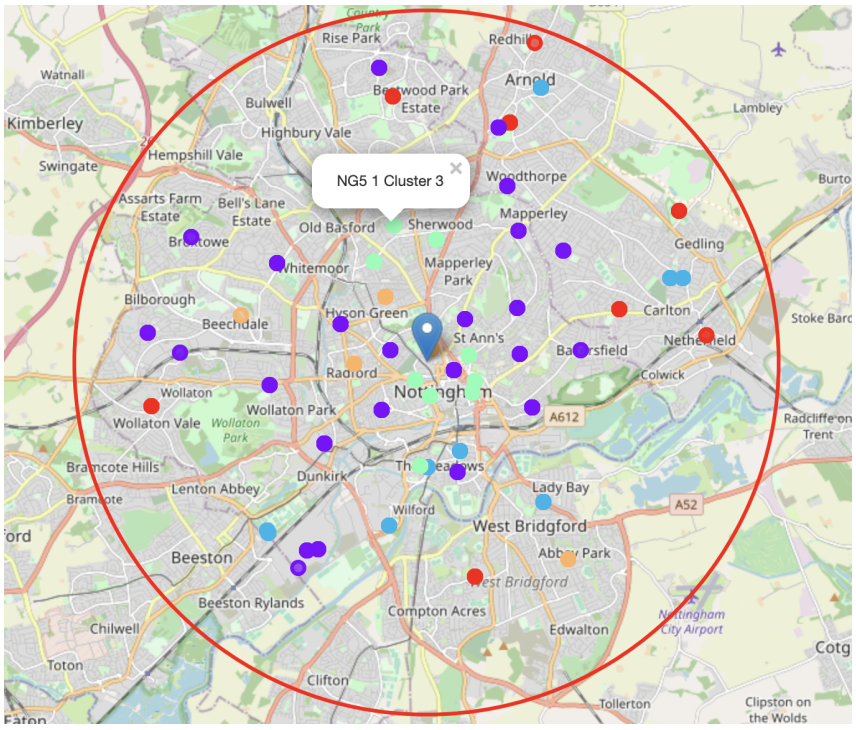

Figure 11: 5 colour coded clusters of Nottingham area around NTU

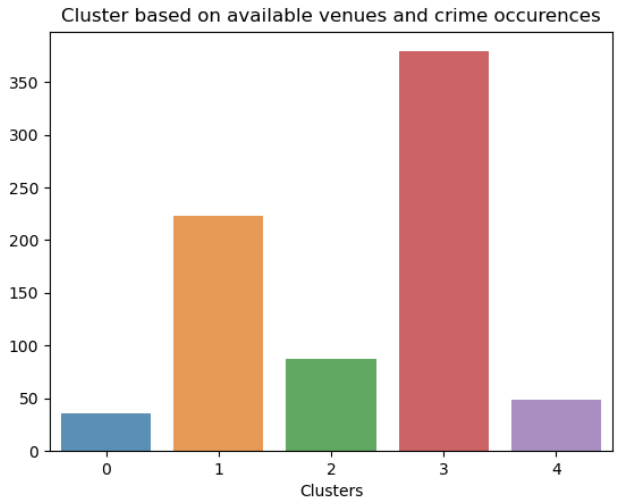

Figure 12: Cluster comparison based on number of venues and crime occurrence

for each location. To find the optimal number of clusters, elbow method was performed. Elbow method shows that the optimal number of k was 5 after which the errors didn't decrease as drastically as shown in Figure 10. I used this k as an optimum number of clusters to cluster areas within 5 km of NTU. The clusters are visualised using folium library. Figure 11 shows colour coded clusters of areas around NTU.

Although, we can visualise clusters, it is not apparent whether one cluster is a better choice than others just on the visual inspection. To compare different clusters, I computed number of venues on each cluster. This provides a good idea about which cluster has good amenities in close proximity i.e within

Area codes in the best two clusters	
Cluster 1	Cluster 3
['NG1 3' 'NG7 2' 'NG7 4' 'NG7 5' 'NG7 1' 'NG3 6' 'NG3 4' 'NG3 3' 'NG3 7' 'NG3 2' 'NG3 5' 'NG2 2' 'NG2 4' 'NG8 1' 'NG8 6' 'NG8 4' 'NG8 9' 'NG8 5' 'NG90 5' 'NG90 4' 'NG90 7' 'NG5 0' 'NG5 9' 'NG5 4']	['NG1 1' 'NG1 9' 'NG1 5' 'NG1 6' 'NG7 7' 'NG3 1' 'NG2 1' 'NG5 2' 'NG5 1']

Table 1: Clusters and area codes within each cluster

500 meters of cluster members. Figure 12 shows that clusters 1 and 3 have significantly higher number of amenities in close proximity. These two clusters are the best options to further investigate into for the best accommodation facilities for students. The postcode areas within each of these clusters are listed in Table 1. These postcode areas would be the best locations within Nottingham around NTU to look into to establish student accommodation business.

5 Discussions and conclusions

In this project, I used cluster analysis of area codes in Nottingham within 5 km distance from Nottingham Trent University to identify ideal places which could potentially be suitable for establishing business providing student accommodation. I used publicly available postal code data to extract location information and I also generated new features for postal code as *Area_code* based on closely situated locations and for crime data. I used crime data to extract number of crimes reported for each location and prepared a new data frame based on this information. These two data frames when combined provided us with good overview of safety of each area. Geospatial maps helped to visualise locations on the map. Another important feature for suitability of the locations is the availability of amenities in close proximity. I used Foursquare API to extract venues around each area and performed cluster analysis based on available amenities and safety of the location. The cluster analysis resulted in 5 different clusters among which two clusters(cluster 1 and cluster 3) showed significantly higher number of amenities. I further listed all the area codes within two of the selected clusters. I propose the solution that these two cluster locations offer the best choices for business owners to attract higher number of students. The proposed locations offer better connectivity and safety.

As a note for further improvements, dividing the crime data further into different categories of crimes, adding transportation facilities and housing costs based on areas can help to improve the recommendation even further.