# **Capstone Project Week 5 IBM Data Science module**

Dr Sofia Korsavva

## **Introduction**

We have been asked to write a python program that will explore the City of Birmingham in the UK for a customer who is interested in purchasing a property for his family. The customer would like to know where in Birmingham he should start looking for properties based on the following requirements:

1. The customer has a budget of between 400-500K in GBP
2. He would like the property to be quite central in the city of Birmingham as him and his family enjoy city life.
3. He would like to know what type of venues exist within 10-15 minutes of the property
4. He would like to know what type of facilities exist within 10-15 of the property, such as supermarkets, other shops and hospitals.
5. The customer is also very worried about the safety of his family so he would like to know in those areas that we are going to recommend to him, the crime rates as officially reported by West Midlands Police, for the past month.

## **Methodology and Data**

To address all the customer requirements we shall develop a python program that will explore some datasets from official sources and will come up with some recommendations for our customer.

In the UK there is an official land registry database that contains information about England and Wales with prices per post code and street name.

The official website is:

<https://www.gov.uk/government/organisations/land-registry>

From there we shall download and process our first csv file that contains all the prices, street names and postcodes for England. We shall process this file as pandas dataframe and we shall extract information for the city of Birmingham, per street name and post code.

From the website <https://www.freemaptools.com/download-uk-postcode-lat-lng.htm>

We shall download the UK csv file that contains all the post codes and maps them to their corresponding latitude and longitudes. We shall then process the file as a pandas framework and extract all the Birmingham post codes that we have identified as per above that we within our customer’s price range. We shall merge the two data sets and map them nicely on a folium map so our customer can have a good visual representation of the information we give him.

We shall then connect to foursquare and extract information for these post codes about venues and facilities in those identified postcodes and present them to our customer so can make an informed decision.

We shall use a k-means clustering algorithm to classify and cluster the selected areas of interest where the customer is now looking to buy property. The customer is mostly interested in the types of facilities around those areas that are not too far away (10-15 minutes). As the customer has a family, he is interested in nearby schools, hospitals, supermarkets etc.

Finally we shall download the csv from the west midlands police website from [http://www.ukcrimestats.com/Police\_Force/West\_Midlands\_Police#](http://www.ukcrimestats.com/Police_Force/West_Midlands_Police)

that contains all the information about the crimes that have occurred in those areas for the past year so our customer can evaluate how safe a particular selection may be.

## **Results**

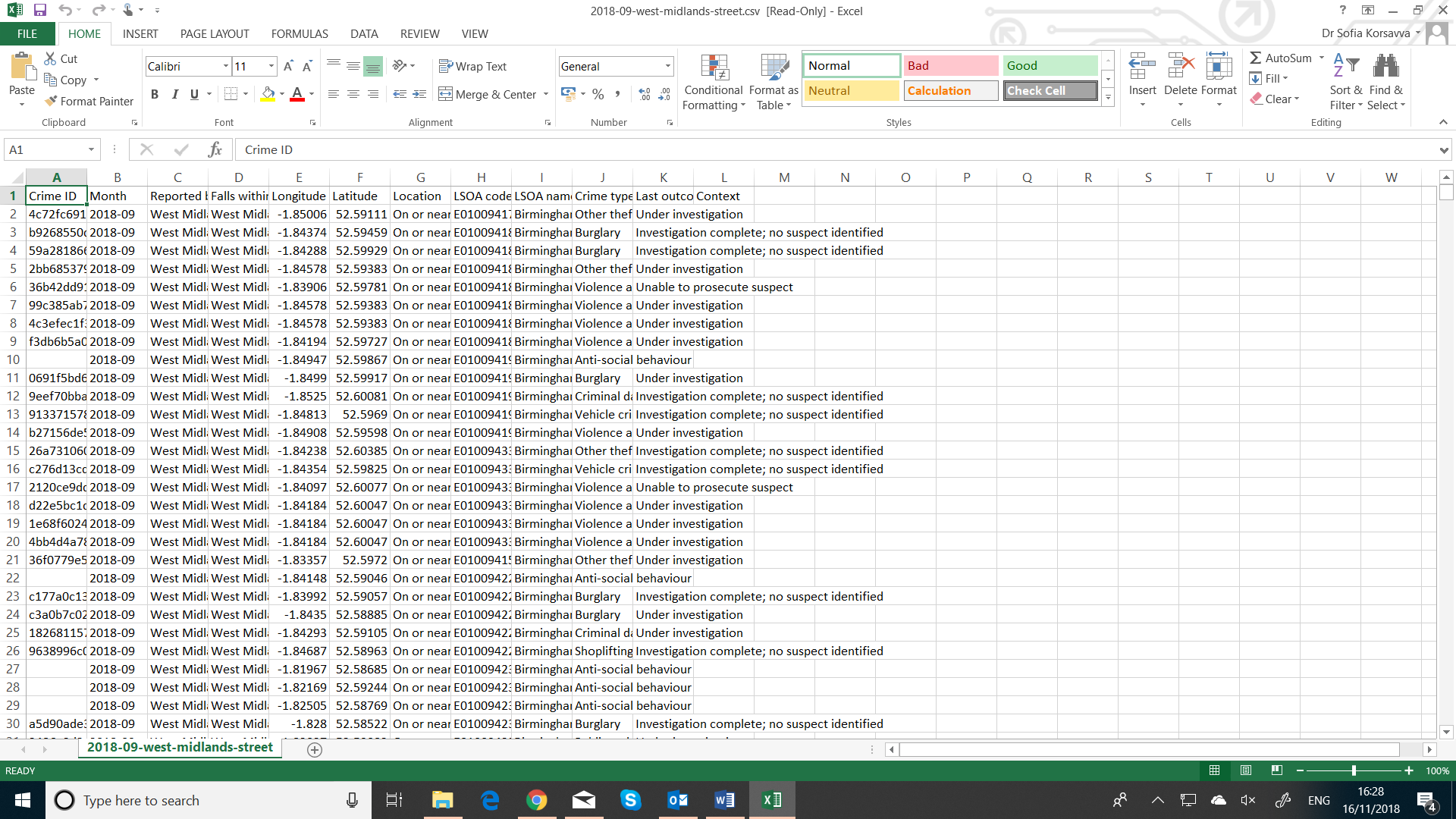
The raw data are all in the form of csv files. As discussed above I have used official UK government websites to obtain house prices per postcode and street names, for the initial data set from the UK land registry database. I include a snapshot of the file which I downloaded below:

A close up of a piece of paper

Description automatically generated

The second csv file I used contained information about UK postcodes, latitudes and longitudes. A snapshot of the file is shown below:

|  |  |  |
| --- | --- | --- |
| Postcode | Latitude | Longitude |
| B1 1AA | 52.47666 | -1.90354 |
| B1 1AD | 52.47666 | -1.90354 |
| B1 1AG | 52.47453 | -1.90216 |
| B1 1AH | 52.47639 | -1.90426 |
| B1 1AQ | 52.47453 | -1.90216 |
| B1 1AT | 52.47639 | -1.90426 |
| B1 1AY | 52.47592 | -1.90539 |
| B1 1AZ | 52.47666 | -1.90354 |
| B1 1BA | 52.47543 | -1.90021 |
| B1 1BB | 52.48078 | -1.9041 |
| B1 1BD | 52.47892 | -1.90294 |
| B1 1BE | 52.47757 | -1.90175 |
| B1 1BF | 52.47489 | -1.90241 |
| B1 1BG | 52.47894 | -1.90315 |
| B1 1BH | 52.47895 | -1.90315 |
| B1 1BJ | 52.47666 | -1.90354 |
| B1 1BL | 52.47685 | -1.90162 |
| B1 1BN | 52.47611 | -1.90072 |
| B1 1BQ | 52.47729 | -1.90278 |
| B1 1BT | 52.47543 | -1.90021 |
| B1 1BU | 52.47542 | -1.89993 |

As the customer wanted also the number of crimes in the selected postcodes that are in his price range I downloaded from the UK police website a csv file that contains all the crime references for all West Midlands and I extracted the crimes in the Birmingham area. The file looks like as per below:



For the venues and facilities information I connected to foursquare and I extracted the information for the preselected areas.

I chose the facilities information and I used the k-means clustering algorithm to cluster the preselected neighbourhoods together. I then presented them on a folium map so my customer could see the neighbourhoods graphically on the map, based on the facilities and conclude about which area would be the best option for him to purchase his property in.

I mainly used pandas dataframes and processed them to collate the information together.

**Results**

The program identified a number of areas that the customer can purchase property based on his requirements.

**Example output**

Postcode Avg\_Price

B12 8LQ 475000

B12 9QA 400000

B13 0AJ 400000

B13 0AL 425000

…. Etc

I also produced the street names and prices for a better understanding of the location.

A screenshot of a social media post

Description automatically generated

In total 40 postcode areas were identified with a price range of 400-500K GBP.

I then merged the postcodes, prices, latitude and longitude together and mapped on the map of Birmingham using folium as per below

A picture containing text, map

Description automatically generated

I then displayed the venues in those areas with the lat and long – I got the venue information from foursquare.

A screenshot of a cell phone

Description automatically generated

I displayed the number of venues in those areas

A screenshot of a cell phone

Description automatically generated

I did the same for the facilities in those areas.

A screenshot of a cell phone

Description automatically generated

I also counted the occurrences of types of facilities in said areas

A screenshot of a cell phone

Description automatically generated

For the facilities I started preparing the output for k-means clustering, starting with hot-encoding:

A close up of a piece of paper

Description automatically generated

I identified the 5 most common facilities

A screenshot of a social media post

Description automatically generated

I then produced the 10 most common facilities in each area

A screenshot of a cell phone

Description automatically generated

I run the k-means clustering algorithm and came up with k labels

A screenshot of a cell phone

Description automatically generated

I then put them on a folium map for my customer, I have clustered the neighbourhoods together based on their similarities in terms of facilities such as schools, hospitals and supermarkets – as the customer has children.

A picture containing text, map

Description automatically generated

Finally I produced a pandas dataframe with the crime reports for the selected areas the customer is looking to buy property. The results were very good, as in those particular areas no crimes had been reported the past month.

**Outcome:** The customer has pricing information in the city of Birmingham UK, the post codes and street names where he can look to purchase properties. He knows the crime rates in those areas, he also has information about venues and facilities. He has a visual representation of the similarity of these areas clustered as per the types of facilities as per his request.

A screenshot of a cell phone

Description automatically generated