House Prices & Neighbourhood Venues Data Analysis of Birmingham, UK

IBM Data Science Coursera Capstone Project Week 4

Introduction

Bermingham is the second-largest city of the United Kingdom (UK) with roughly 1.1 million inhabitants within the city area and 3.8 million inhabitants within the metropolitan area. This also makes Birmingham the 17th largest city in the world. Located approximately 100 miles from central London, Birmingham, as one of the United Kingdom's major cities, is considered to be the social, cultural, financial, and commercial center of both the East and West Midlands.

There are some facts about the city

1. **Population:** 1,111,307 (2017 estimate), 3.8 million in the Greater Metropolitan area, most ethnically diverse city in the UK.
2. **Official Language:** English (*de facto)* with 108 languages being spoken in schools, including Urdu, Punjabi, Bengali, and many others.
3. **GDP Per Capita:** $46,296.98 USD
4. **Youthful Population:** 45.7% of Birmingham's population is under 30
5. **Currency:** Pound Sterling or Great Britain Pound (£, GBP). As of early 2017, $1 USD = £0.80, €1 EUR = £0.85, $1 CAD = £0.61, $1 AUD = £0.62

Compared with other major cities in the world, houses and flats and flat prices in Birmingham are reasonable. From a new person's point of view, if they want to move to this city they want to invest in such places where the housing prices are low and the facilities(shops, restaurants, parks, hotels, etc.) and social venues are nearby.

Keeping these things into consideration it is very difficult for an individual to find such a place in such a large city and gather this much information. To solve these problems, we can create a map and information chart where the real estate index is placed on the city and each area is clustered according to the venue density.

Data Collection

There are various sources are used to collect the data for this exercise -

To find all the areas and postcode district in and around the city

https://en.wikipedia.org/wiki/B_postcode_area

To find the real estate details (Average price etc) following service is used as the data source
https://propertydata.co.uk/cities/Birmingham

https://opencagedata.com/ for getting geo coordinates for the areas.

Forsquare API to get the most common venues of Bermingham

Data Preprocessing

Data is collected from Wikipedia for the list of area codes and details. City area details are collected after parsing HTML DOM via Python.

Postcode district ⇕	Post town ⇕	Coverage ⇕	Local authority area ⇕
B1	BIRMINGHAM	Birmingham City Centre, Broad Street (east)	Birmingham
B2	BIRMINGHAM	Birmingham City Centre, New Street	Birmingham
B3	BIRMINGHAM	Birmingham City Centre, Newhall Street	Birmingham
B4	BIRMINGHAM	Birmingham City Centre, Corporation Street (north)	Birmingham
B5	BIRMINGHAM	Digbeth, Highgate, Lee Bank	Birmingham
B6	BIRMINGHAM	Aston, Witton	Birmingham
B7	BIRMINGHAM	Nechells	Birmingham
B8	BIRMINGHAM	Washwood Heath, Ward End, Saltley	Birmingham
B9	BIRMINGHAM	Bordesley Green, Bordesley	Birmingham
B10	BIRMINGHAM	Small Heath	Birmingham
B11	BIRMINGHAM	Sparkhill, Sparkbrook, Tyseley	Birmingham
B12	BIRMINGHAM	Balsall Heath, Sparkbrook, Highgate	Birmingham
B13	BIRMINGHAM	Moseley, Billesley	Birmingham
B14	BIRMINGHAM	Kings Heath, Yardley Wood, Druids Heath, Highter's Heath, Warstock	Birmingham
B15	BIRMINGHAM	Edgbaston,	Birmingham
B16	BIRMINGHAM	Ladywood	Birmingham
B17	BIRMINGHAM	Harborne, Edgbaston	Birmingham

HTML DOM of this page consists of a table and few hyperlinks. So, after collecting HTML DOM table , collecting data is converted into dataframe. And new dataframe looked like -

	Town	Coverage	Area	Postal
1	BIRMINGHAM	Birmingham City Centre, Broad Street (east)	Birmingham	B1
2	BIRMINGHAM	Birmingham City Centre, New Street	Birmingham	B2
3	BIRMINGHAM	Birmingham City Centre, Newhall Street	Birmingham	B3
4	BIRMINGHAM	Birmingham City Centre, Corporation Street (no...	Birmingham	B4
5	BIRMINGHAM	Digbeth, Highgate, Lee Bank	Birmingham	B5

After collecting the data of Bermingham areas (coverage), we need to collect the average price of these areas. For that price data from is used from PropertyData (https://propertydata.co.uk/cities/Birmingham) .

Data is read from the URL PropertyData. HTML DOM of PropertyData is parsed and property price data is collected like -

|	Postal	AvgYield	AvgPrice	PerSqftPrice	FiveYearChange	Links
1	B1	4.8%	£218,216	£317	+19%	Explore data
2	B3	4.3%	£239,971	£286	+23%	Explore data
3	B5	5.2%	£211,124	£258	+19%	Explore data
4	B9	5.3%	£139,358	£134	+23%	Explore data
5	B11	4.0%	£171,299	£143	+22%	Explore data

A data cleaning needs to be done here. From the table above we only require Postal and AvgPrice. So a new data set is created merging these two -

	Town	Coverage	Area	Postal	AvgPrice
0	BIRMINGHAM	Birmingham City Centre, Broad Street (east)	Birmingham	B1	218216
1	BIRMINGHAM	Birmingham City Centre, Newhall Street	Birmingham	B3	239971
2	BIRMINGHAM	Digbeth, Highgate, Lee Bank	Birmingham	B5	211124
3	BIRMINGHAM	Bordesley Green, Bordesley	Birmingham	B9	139358
4	BIRMINGHAM	Sparkhill, Sparkbrook, Tyseley	Birmingham	B11	171299

Another challenge is to find the geolocation (latitude and longitude). For that Opencagedata service is used.

	Town	Coverage	Area	Postal	AvgPrice	lat	lng
0	BIRMINGHAM	Birmingham City Centre, Broad Street (east)	Birmingham	B1	218216	52.478955	-1.905994
1	BIRMINGHAM	Birmingham City Centre, Newhall Street	Birmingham	B3	239971	52.480859	-1.900922
2	BIRMINGHAM	Digbeth, Highgate, Lee Bank	Birmingham	B5	211124	52.886993	-2.886859
3	BIRMINGHAM	Bordesley Green, Bordesley	Birmingham	B9	139358	52.479293	-1.838810
4	BIRMINGHAM	Sparkhill, Sparkbrook, Tyseley	Birmingham	B11	171299	52.463967	-1.875828

At this point we have all the details required like Town, Coverage, Area and GeoDetails (Lat, Lng). To represent these geolocations and area on a map we use Folium. Folium is a python package to show the map. Folium is used to visualize geographic details of Birmingham and its coverage. All the latitude and longitude are passed to the Folium map method to achieve following visual -
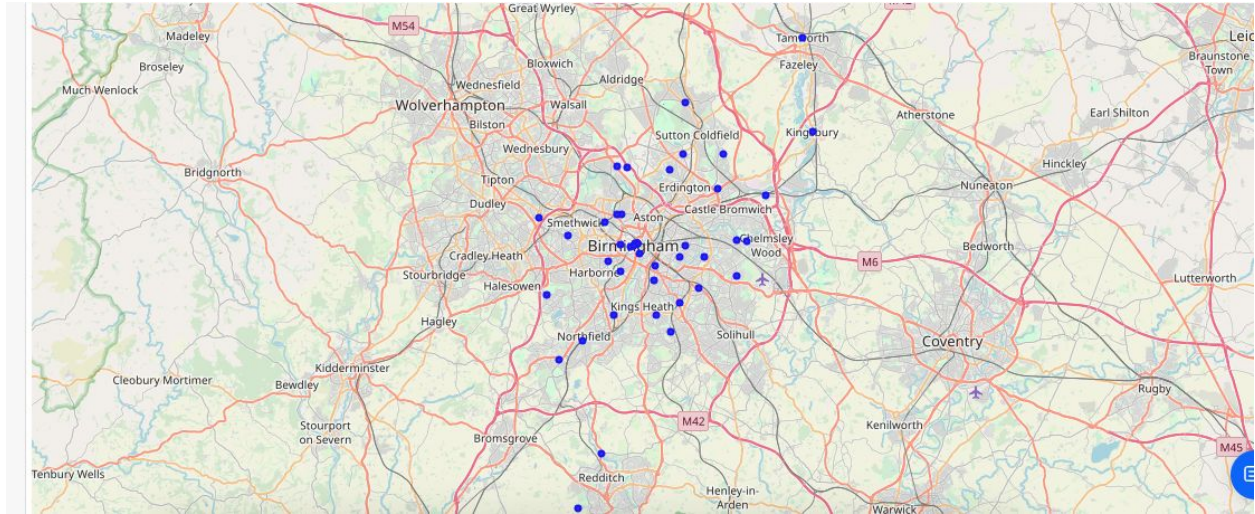
We use Foursquare API to explore the coverage and segment them. For our project, we designed the limit as 100 venue and the radius 1400 meter for each borough from their given latitude and longitude information. Here is the head of the list Venues name, category, latitude and longitude information from Foursquare API.