Coursera Capstone Project : Applied Data Science

Pratik Saha

SRMIST, Chennai, India

pratiksaha198@gmail.com

Overview

- Introduction
- Business Problem
- Data
  - Neighbourhoods
  - Geocoding
  - Venue Data
- Methodology
  - Accuracy of the Geocoding API
  - Folium
  - One hot encoding
  - Top 10 most common venues
  - Optimal number of clusters
  - K-means clustering
- Results
- Discussion
- Conclusion

Introduction :

For the last decade, the United Kingdom’s grocery landscape has been dominated by the ‘big four’ supermarket chains: Tesco, Asda, Sainsbury’s and Morrisons. However, on the back of the economic recession, rising food prices and tightened belts, the market has been shaken by British consumers' search for value.

In 2015, IGD valued the grocery retail market at 178 billion British pounds, predicting this figure to rise annually up to 2021. Additional figures from the Office of National Statistics show that despite an increase in the value of grocery store sales , the volume of goods purchased by consumers shows negligible change.

Although rising food prices have not caused the quantity of goods purchased to fall, consumers seem more likely than ever to search for cheaper alternatives to the ‘big four.’ Discount supermarkets are enjoying a surge in popularity among food shoppers. According to figures from Kantar Worldpanel, all of the leading four supermarket brands have lost market share in the three months to August 2016.

Therefore the greater accessibility to these areas fueled by long term growth plans are necessary for the revival of these supermarkets along with greater development of the neighboring areas.

Business Problem :

Consumer surveys indicate a shift in thinking among shoppers. According to recent evaluations, the number of consumers that use discounters is still increasing, while the number that never use them has decreased: they are chosen over supermarkets because of their public perception as cheaper and, ironically, to avoid the complexity of over-promotion. Meanwhile, online grocery shopping could further revolutionize the market as e-commerce gains popularity among shoppers in the United Kingdom. Currently, online grocery sales in the United Kingdom take 6.9 percent of the global e-commerce market, and is thus the largest online grocery market in Europe.

Thus in hope of the supermarkets to keep up their revenue , the main objective of this project is to find the best locations to situate the supermarkets so that it has greater accessibility of buyers and draws more attention along with keeping in mind the surrounding neighborhoods so that that area also indirectly thrives due to the supermarkets like food
joints , movie-halls , etc.

Data :

Neighborhoods :

The data of the neighbourhoods in Greater Manchester can be extracted out by web scraping using BeautifulSoup library for Python. The neighbourhood data is scraped from a Wikipedia webpage.

Geocoding :

The file contents are retrieved into a Pandas DataFrame. The latitude and longitude of the neighbourhoods are retrieved using OpenCage Geocoding API. The geometric location values are then stored into the initial dataframe.

Venue Data:

From the location data obtained after Web Scraping and Geocoding, the venue data is found out by passing in the required parameters to the FourSquare API, and creating another DataFrame to contain all the venue details along with the respective neighbourhoods.

Methodology :

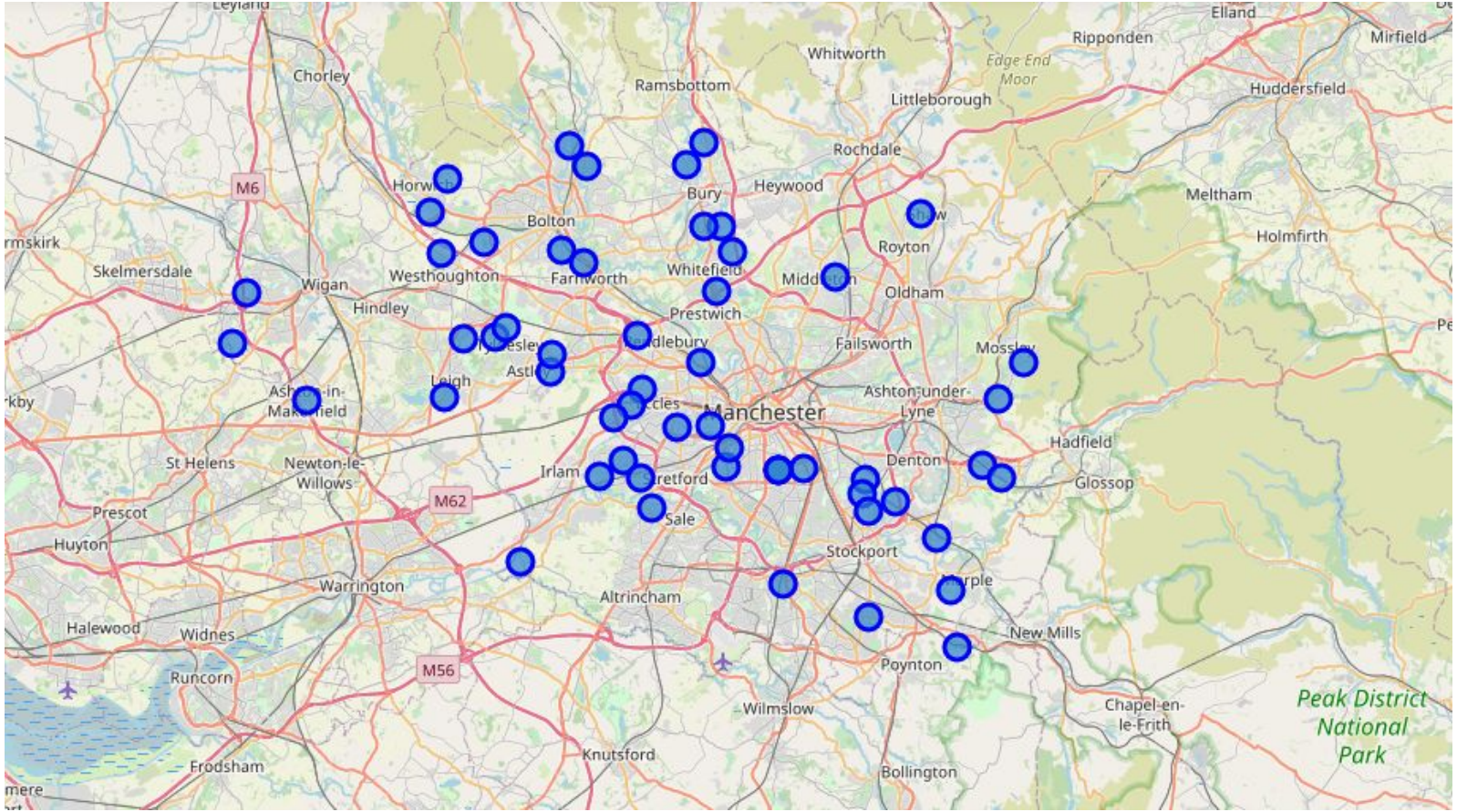
Accuracy of Geocoding API

In the initial development phase with Google Maps Geocoder API, the number of erroneous results were of an appreciable amount, which led to the development of an algorithm to analyze the accuracy of the Geocoding API used. In the algorithm developed, Geocoding API from various providers were tested, and in the end, OpenCage Geocoder API turned out to have the least number of collisions.

Folium

Folium builds on the data wrangling strengths of the Python ecosystem and the mapping strengths of the leaflet.js library. All cluster visualizations are done with help of Folium which in turn generates a Leaflet map made using OpenStreetMap technology.

Neighborhoods of Greater Manchester Area :



One hot encoding

One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. For the K-means Clustering Algorithm, all unique items under Venue Category are one-hot encoded.

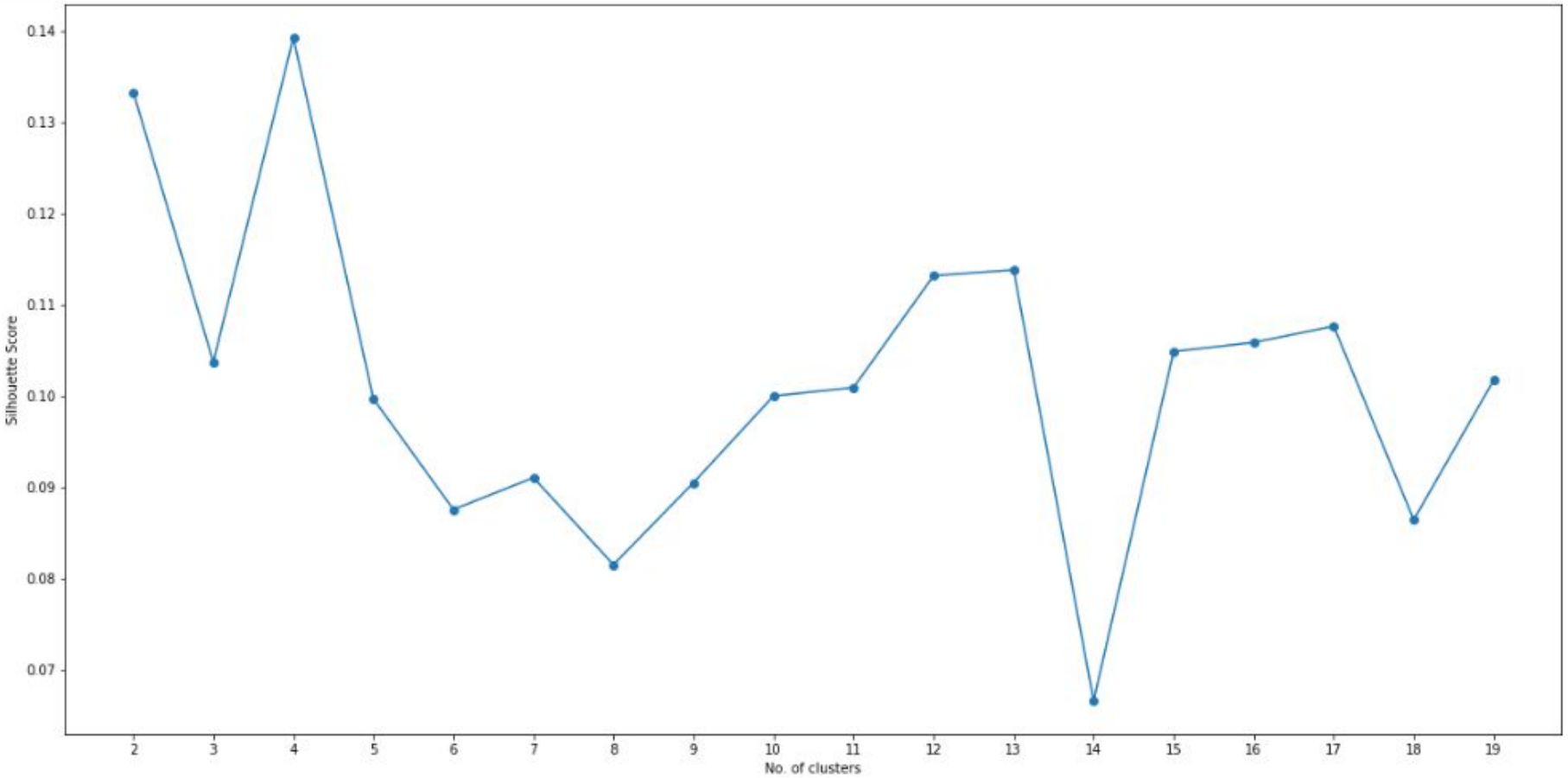
Top 10 most common venues

Due to high variety in the venues, only the top 10 common venues are selected and a new DataFrame is made, which is used to train the K-means Clustering Algorithm.

Optimal number of clusters

Silhouette Score is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from -1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. Based on the Silhouette Score of various clusters below 20, the optimal cluster size is determined.

Finding the optimal k value :

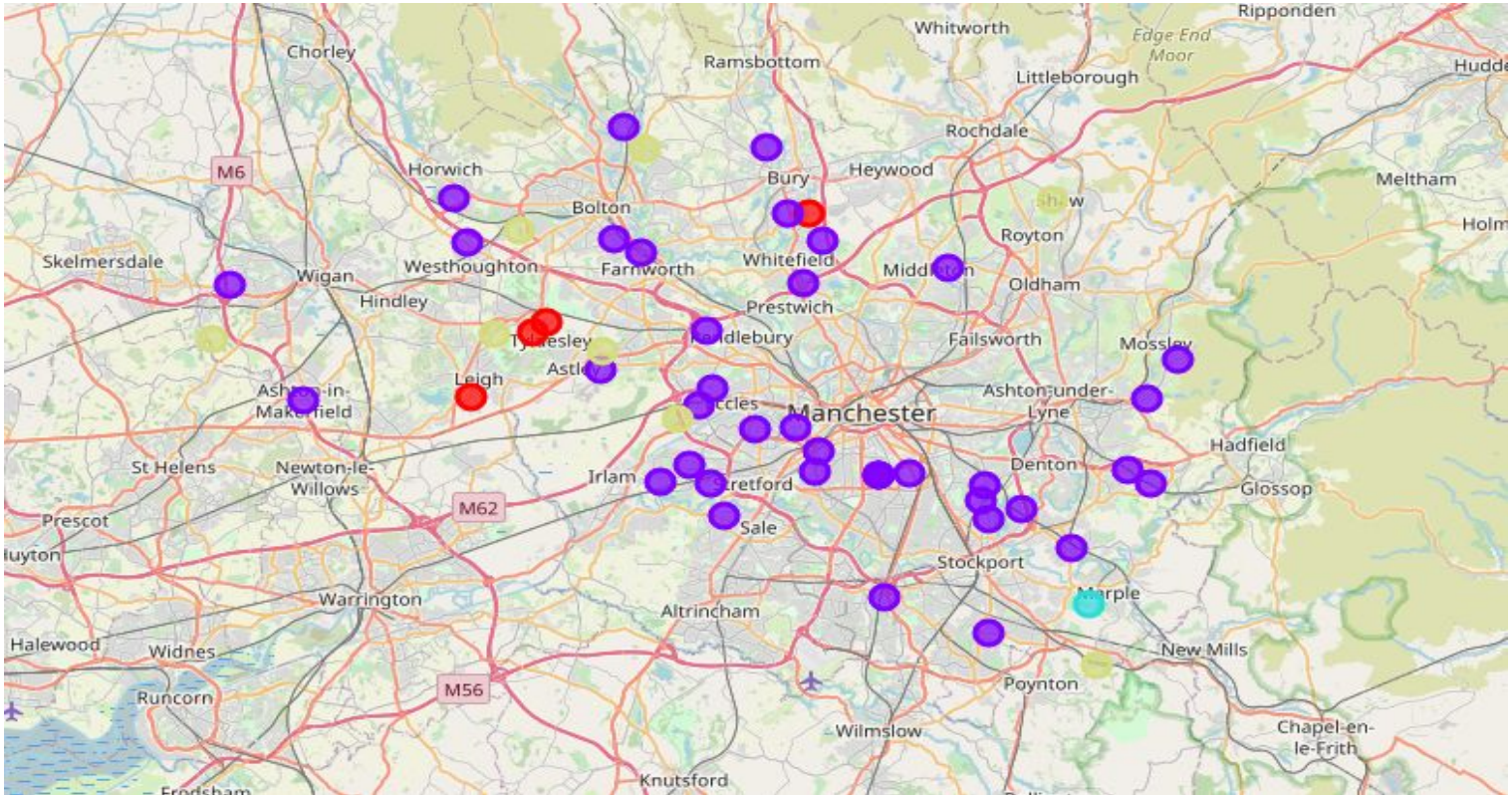


K-means clustering

The venue data is then trained using K-means Clustering Algorithm to get the desired clusters to base the analysis on. K-means was chosen as the variables (Venue Categories) are huge, and in such situations K-means will be computationally faster than other clustering algorithms.

Results :

The neighbourhoods are divided into n clusters where n is the number of clusters found using the optimal approach. The clustered neighbourhoods are visualized using different colours so as to make them distinguishable.



Discussion :

After analyzing the various clusters produced by the Machine learning algorithm, cluster no. 2 , is a prime fit to solving the problem of finding a cluster with a common venue as a train station mentioned before.

	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
16	Hindsford	Supermarket	Bar	Roller Rink	Fast Food Restaurant	Soccer Field	Ethiopian Restaurant	Entertainment Service	Event Space	Falafel Restaurant	Fried Chicken Joint
34	Pennington, Greater Manchester	Supermarket	Gym	Movie Theater	Chinese Restaurant	Portuguese Restaurant	Hotel	Sports Club	Stadium	Gastropub	Playground
35	Pilsworth	Supermarket	Gas Station	Hotel	Turkish Restaurant	Fish Market	Fish & Chips Shop	Film Studio	Fast Food Restaurant	Farm	Falafel Restaurant
43	Shakerley	Supermarket	Pub	Train Station	Fast Food Restaurant	Fish Market	Fish & Chips Shop	Film Studio	Farm	Falafel Restaurant	Turkish Restaurant

These four places Hindsford , Pennington , Pilsworth , Shakerley fall in the heart of the Greater Manchester area. This is in direct correlation to them having the highest footfall and daily agglomeration of people of all works of life having different jobs and places to visit.

Conclusion :

The four identified places in Greater Manchester , United Kingdom have the potential for supermarkets to thrive if developed there. This is due to them being one of the most central and most accessible locations in Manchester having great railway connectivity and bus transport capability. Also these four areas have a large number of small localized markets of different aspects like movie-theaters , gyms , hotels and restaurants which can profit indirectly due to the increased footfall in these areas if supermarkets are localised in these places.