Coursera Capstone Project Applied Data Science

Author : Vishwajeet Patil

1. 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.

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.

2. Business Problem

Consumer surveys indicate a shift in thinking among shoppers. According to recent evaluations .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. 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 neighbourhoods.

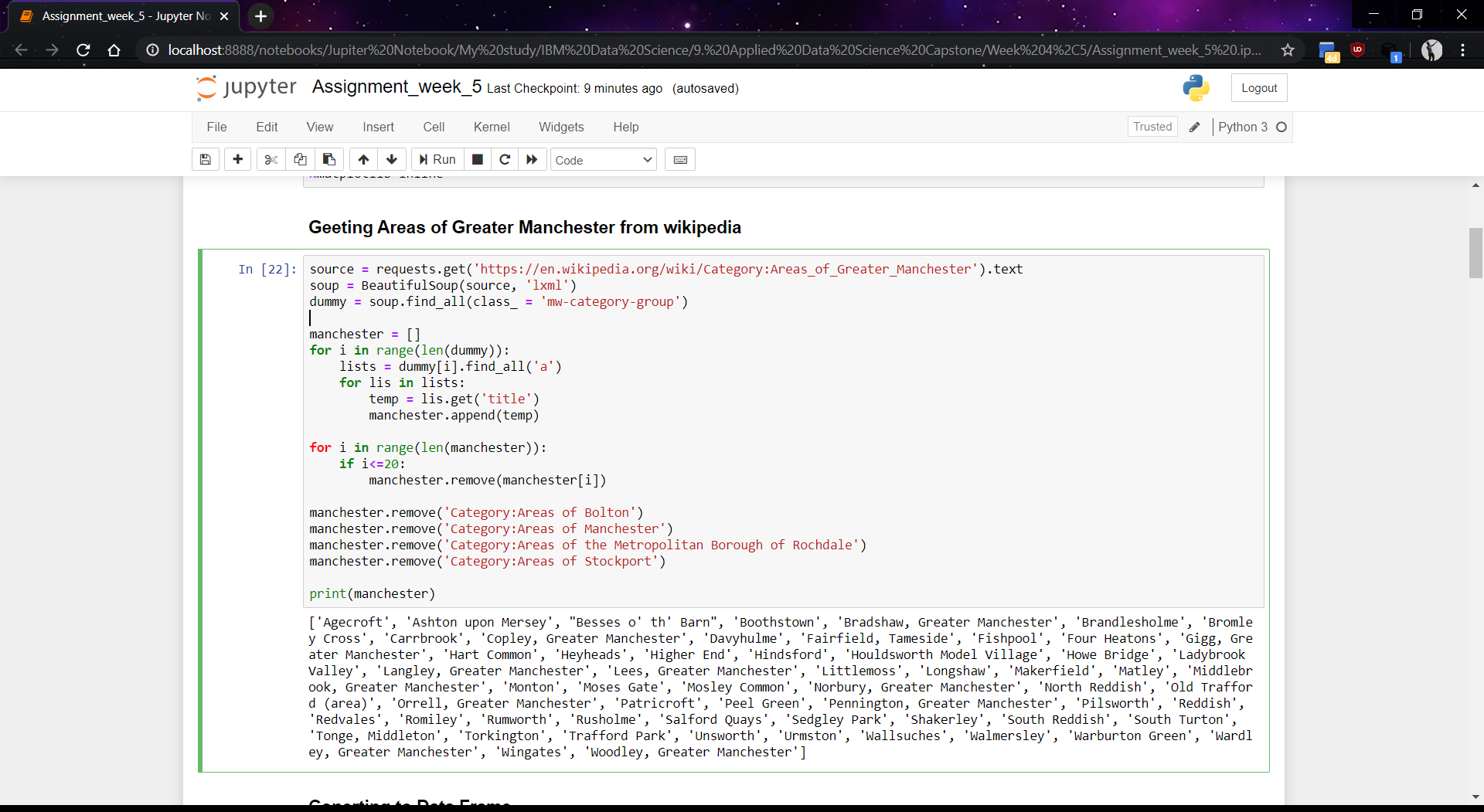
3. Data :-

The data for this project has been collected from multiple sources.

3.1 Neighbourhoods :-

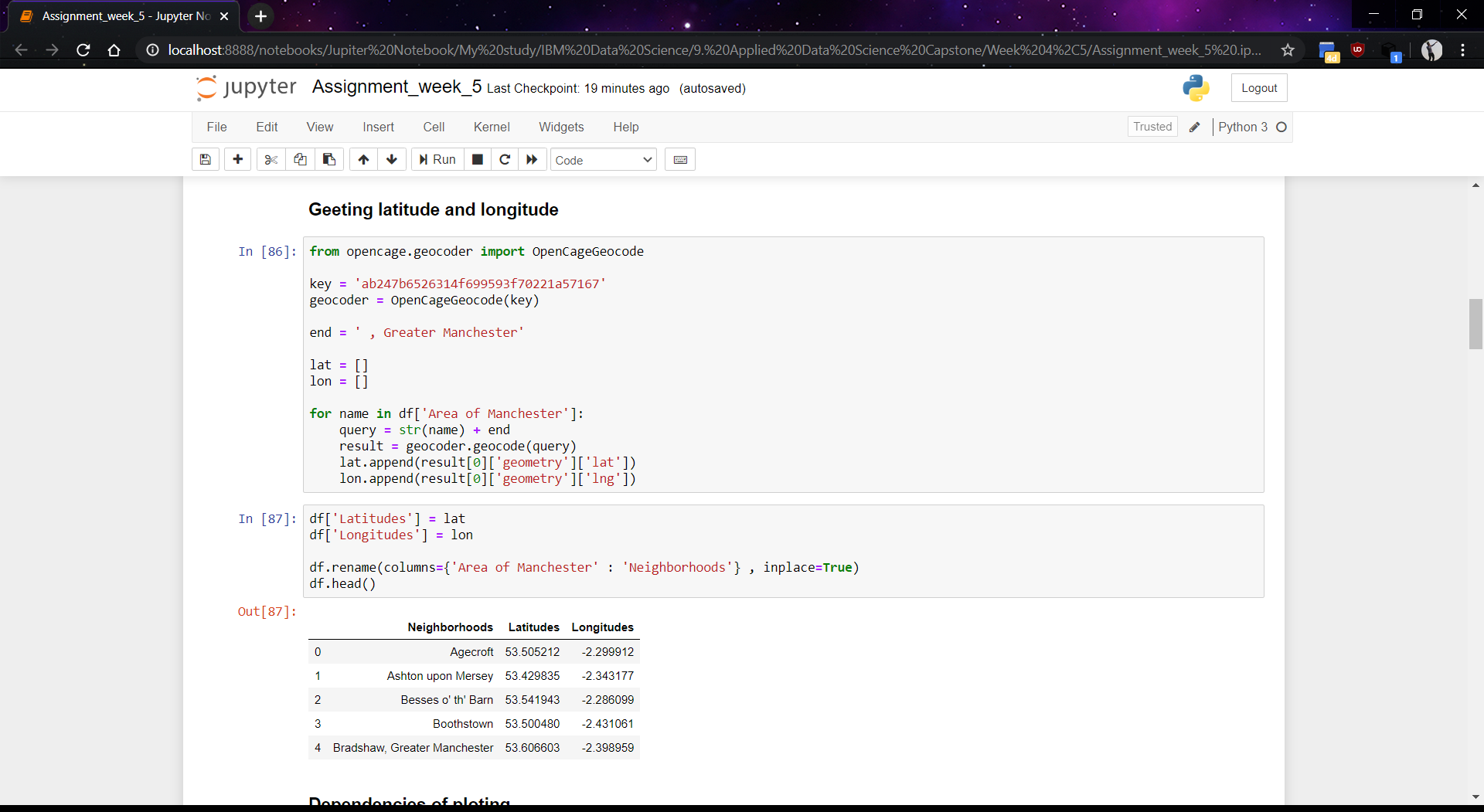
The data of the neighbourhoods in Greater Manchester is extracted out by web scraping using BeautifulSoup library for Python. The neighbourhood data is scraped from a Wikipedia

url=’ <https://en.wikipedia.org/wiki/Category:Areas_of_Greater_Manchester>’



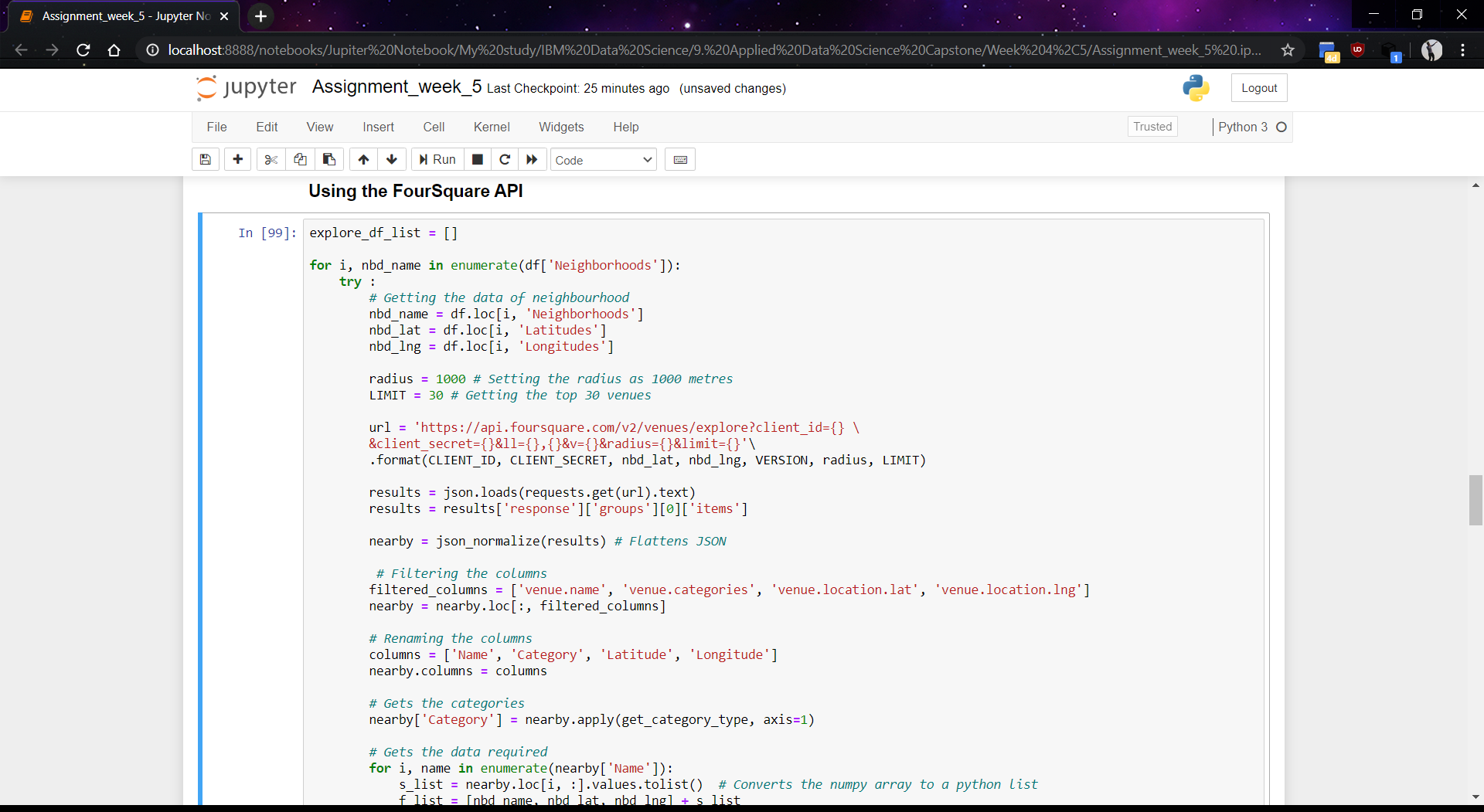
3.2 Geocoding :-

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.



3.3 Venue Data :-

FourSquare API is used to find the venue data and new dataframe is created along with the respective neighbourhoods.

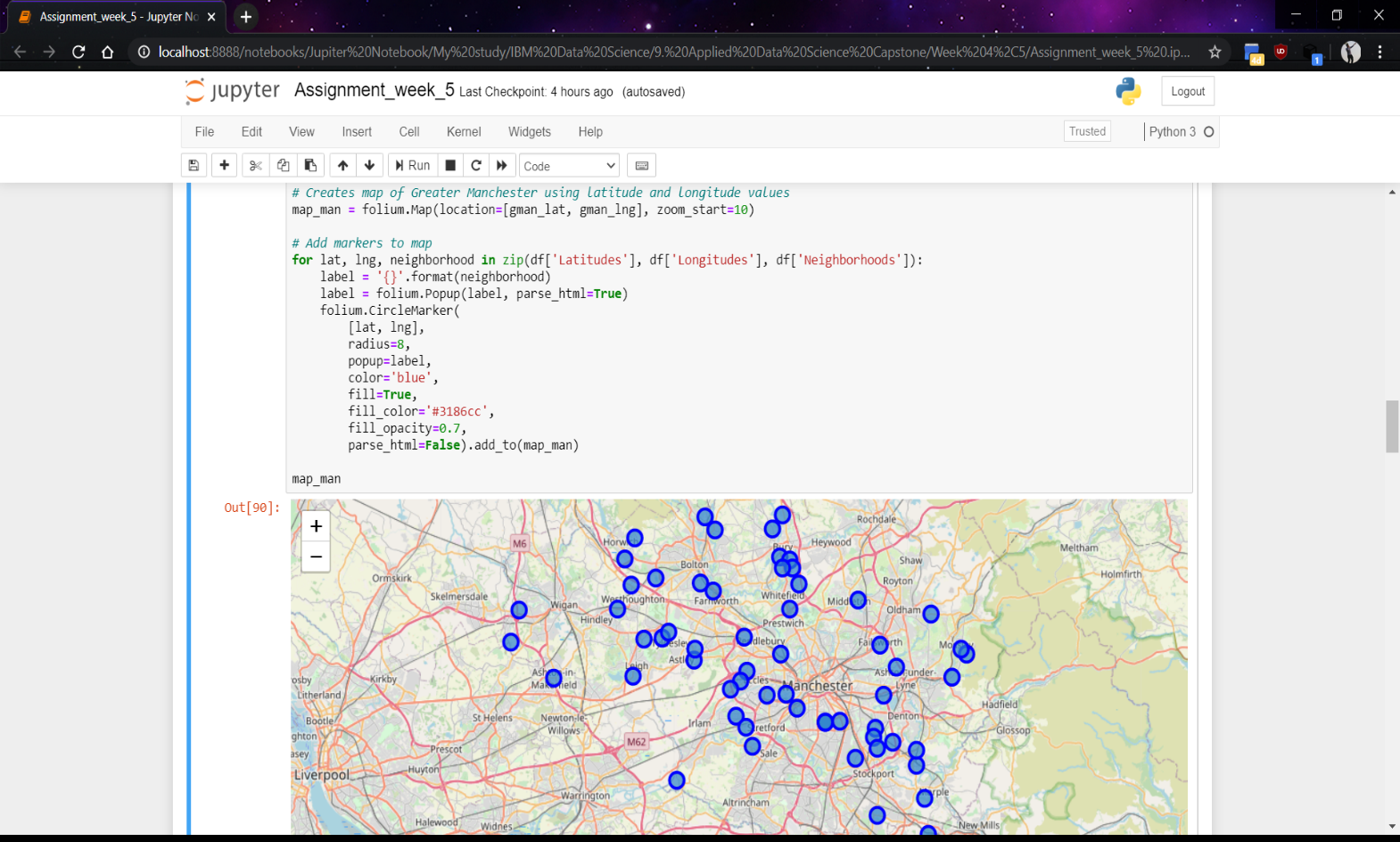


4. Methodology :-

Different methods and algorithms are use to check the data and make sure the prediction are accurate and worthy.

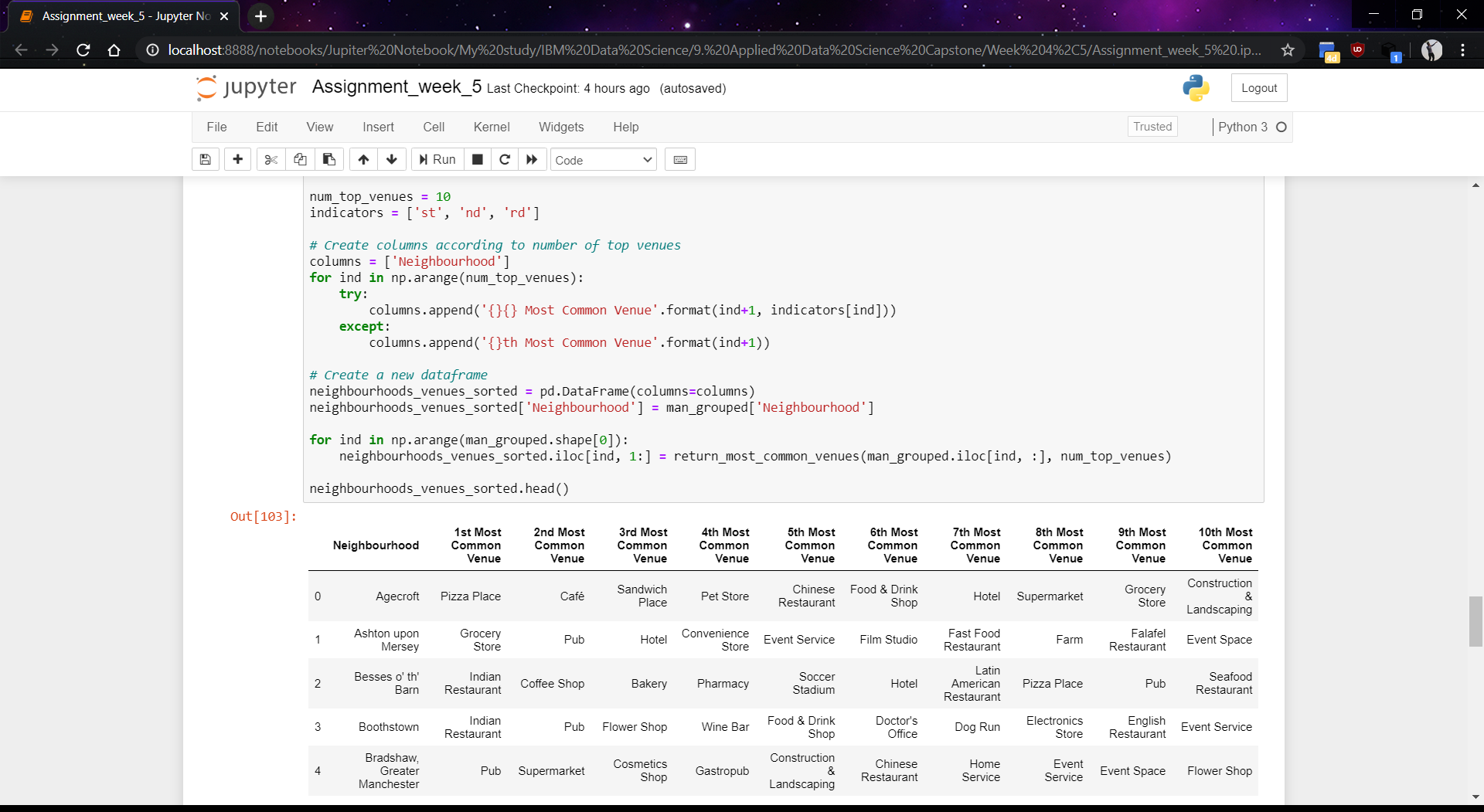
4.1 Folium :-

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.



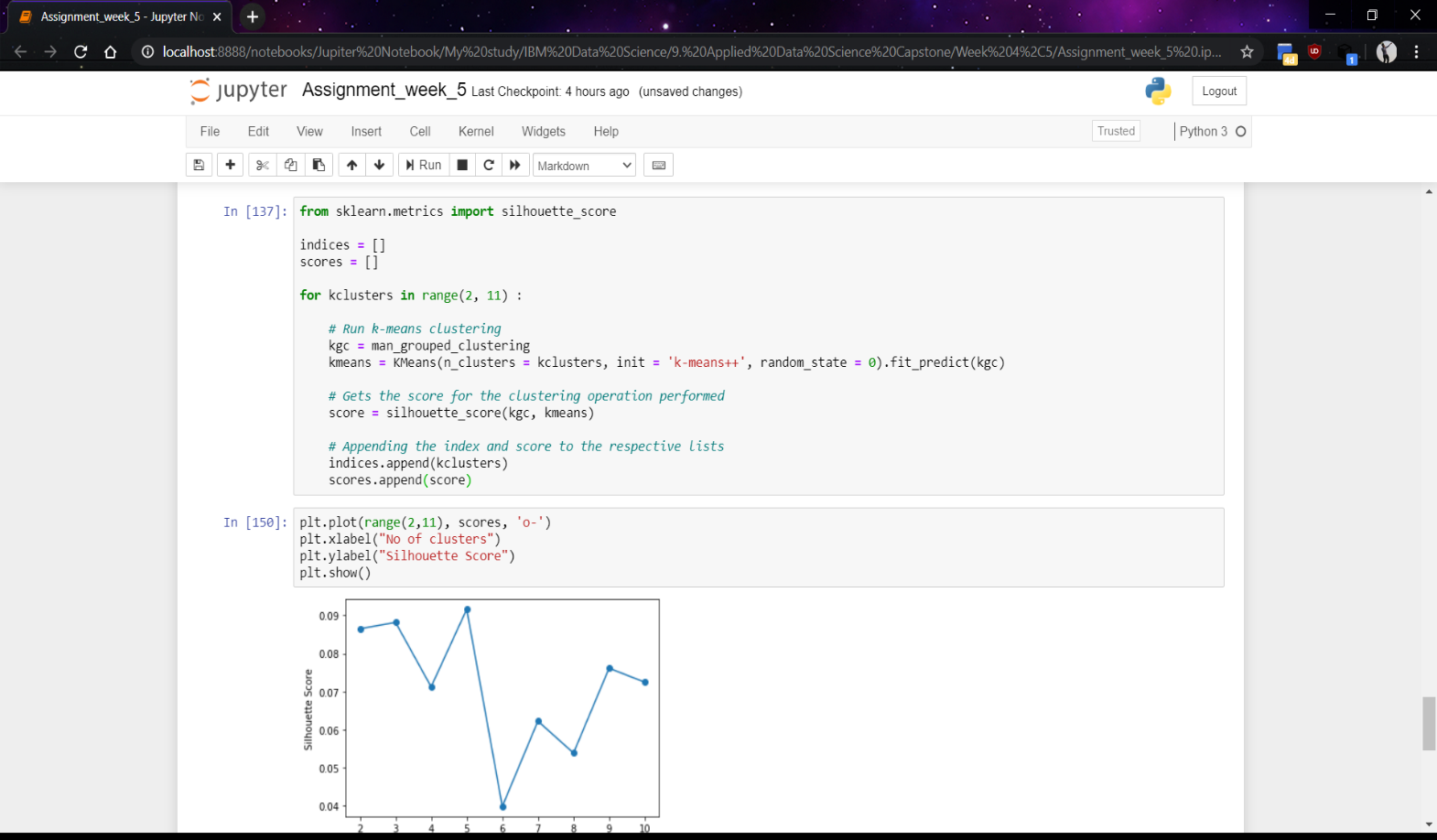
4.2 Top 10 most common venues :-

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.



4.3 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).

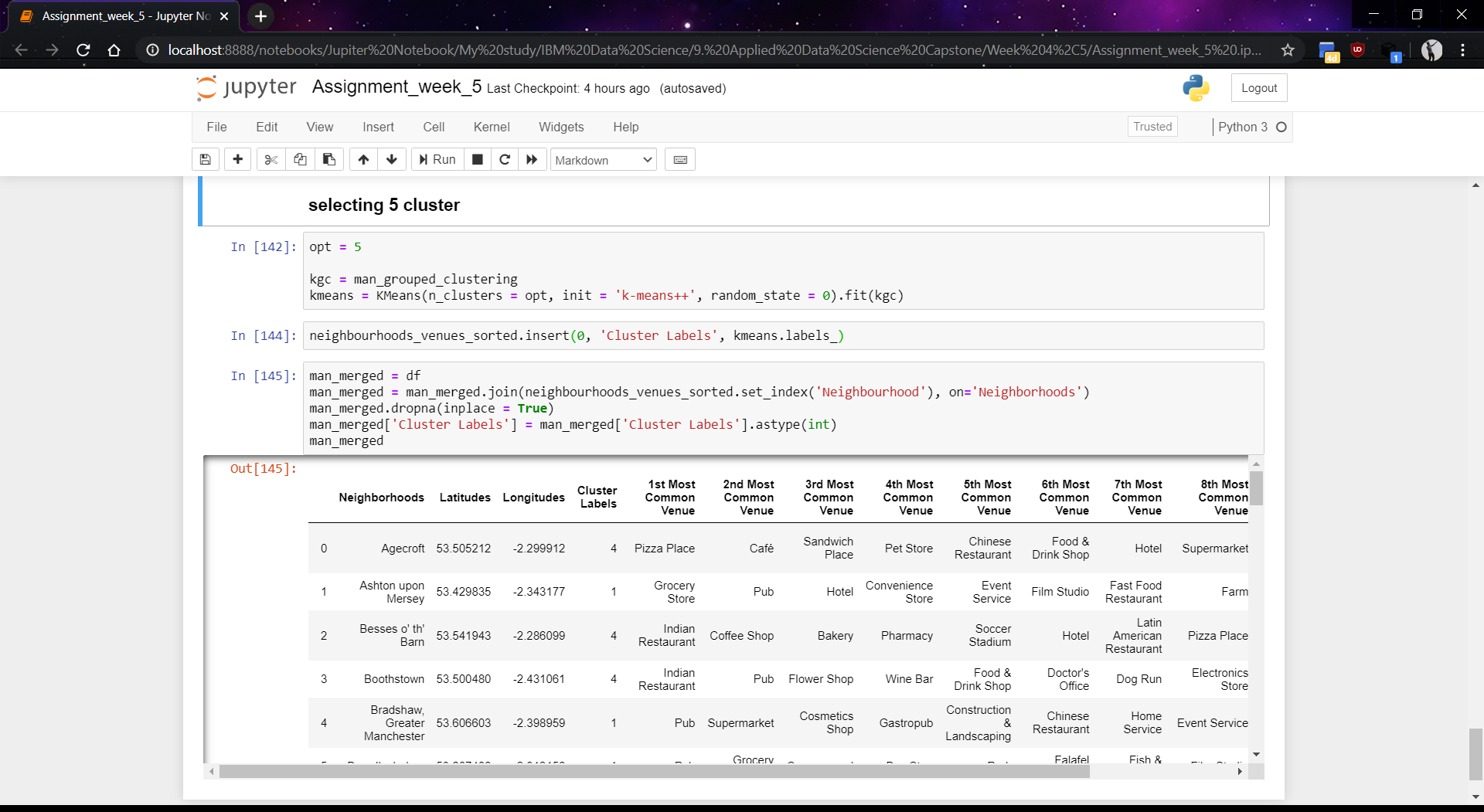


4.4 K-means clustering :-

The venue data is then trained using K-means Clustering Algorithm to get the desired clusters to base the analysis on.

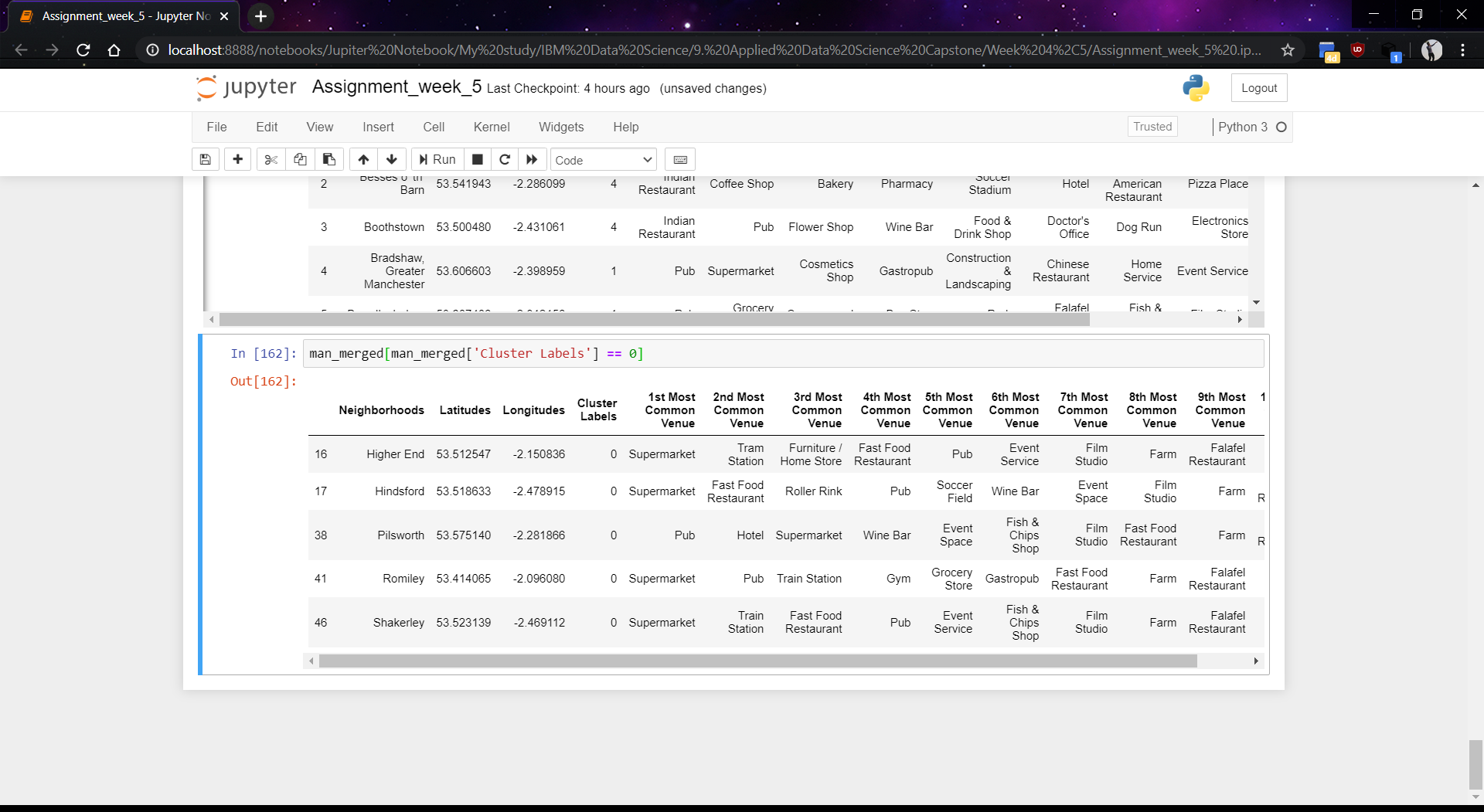
5. Results :-

The neighbourhoods are divided into n (5 here) clusters where n is the number of clusters found using the optimal approach.



6 Discussion :-

After analyzing the various clusters produced by the Machine learning algorithm, cluster no. 0 , is a prime fit to solving the problem of finding a cluster with a common venue as a train station mentioned before.



7 Conclusion :-

These four places Hindsford , Romiley, Pilsworth , Shakerley fall in the heart of the Greater Manchester area. 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.