Coursera Capstone

IBM Applied Data Science Capstone

1. To know the best place to start living in the city of London

2. To start a bar in the city of London .

By: Mohammad Aslam Mahmood

July 2020.

Introduction

Every individual dreams of starting living in a neighbourhood with a lot of considerations such as proximity to the workplace,public transport availability ,crime rate and many more.

The important among all of these factors are happiness_index,life_satisfaction_score.

Secondly,

A bar is a high in demand place all over the world.The annual global average **alcohol consumption** is 6.4 liters per person older than 15 (in 2016). To account for the differences in **alcohol** content of different **alcoholic** drinks (e.g. beer, wine, spirits), this is reported in liters of pure **alcohol** per year.

Business Problem

1)The objective of this capstone project is to analyse and select the best locations in the city of London to start a family or living or moving in from other places

2).The best place in the city of london to start a "Bar".

Using the techniques learnt throughout the course I aim to solve the problem.

Target Audience of this project

Housing corporations,police,hospitals,yoga centre owners and many more businesses as business flourishes where people are happy and satisfied .---"general business class people "

Bar owners looking to start or expand their stores in the city of london.

Data

To solve the problem, we will need the following data:

1. • List of neighbourhoods in London. This defines the scope of this project which is

   confined to the city of London • Latitude and longitude coordinates of those neighbourhoods are required in order to plot the map and get the venue data.

2. London borough profile data for the living indices and area names .

Sources of data and methods to extract them

https://data.london.gov.uk/download/london-borough-profiles/c1693b82-68b1
-44ee-beb2-3decf17dc1f8/london-borough-profiles.csv"

The data is available on the london datastore and the link above gives us what we need .It can be extracted using pandas read csv method .

Then we will get the geographical coordinates of the neighbourhoods using

Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods

We will divide the regions into clusters and decide on the best place to live.

After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare

has one of the largest databases of 105+ million places and is used by over 125,000 developers.

Foursquare API will provide many categories of the venue data, we are particularly interested in the

Bar category in order to help us to solve the business problem put forward. This is a

project that will make use of many data science skills, from web scraping (Wikipedia), working with

API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map

visualization (Folium). In the next section, we will present the Methodology section where we will

discuss the steps taken in this project, the data analysis that we did and the machine learning

technique that was used.

Methodology

Firstly, we need to get the list of neighbourhoods in the city of london.
https://data.london.gov.uk/download/london-borough-profiles/c1693b82-68b1-44ee-beb2-3decf17dc1f8/london-borough-profiles.csv"

The data is available on the london datastore and the link above gives us what we need .It can be extracted using pandas read csv method . We need to get the geographical

coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using the Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of London

Now we will cluster the data of the happiness indices of the city of the london and map using folium ,similarly ,

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key.

We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analysing the "Bar" data, we will filter the "Bar" as a venue category for the neighbourhoods. Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as

possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project.

 We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for "Bar". The results will allow us to identify which neighbourhoods have higher concentration of Bar while neighbourhoods have fewer numbers of bars.

Based on the occurrence of shopping malls in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new shopping malls.

Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the living index
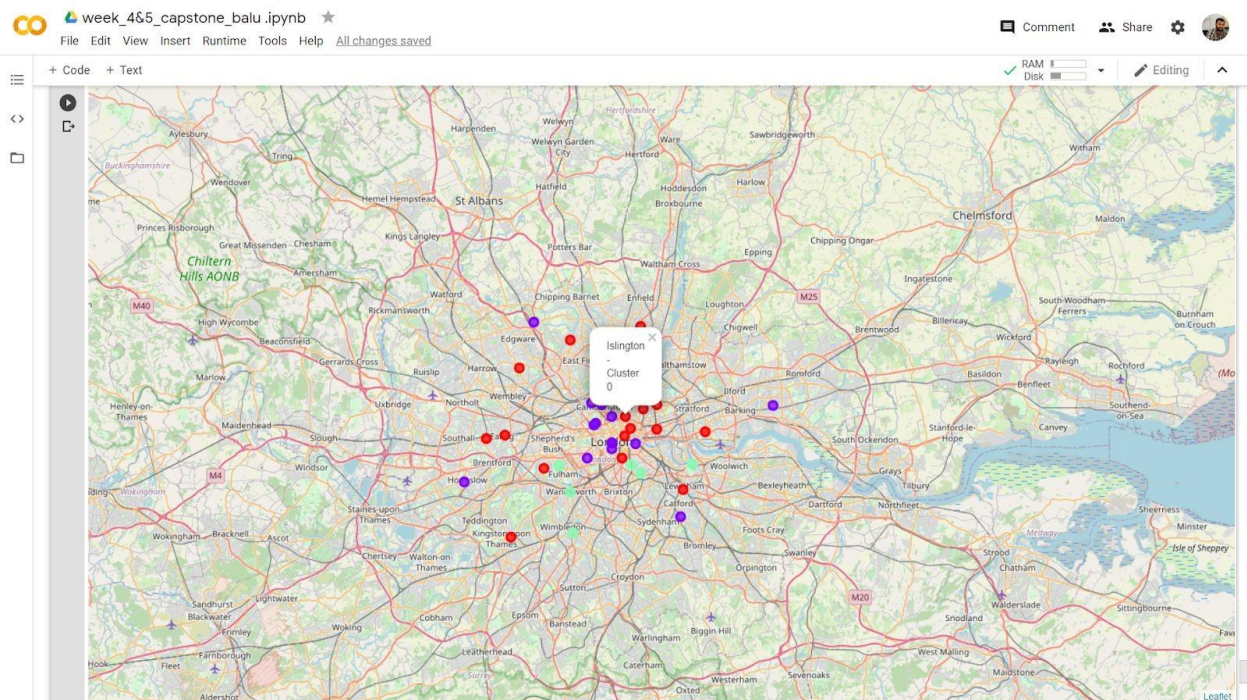
All the areas in the city of London ,England have been clustered into three regions as shown above in the map .They have been clustered based on the happiness index,worthwhileness index ,life satisfaction indices that give the overall living index of these areas .

The results of the clustering are visualized in the maps below.
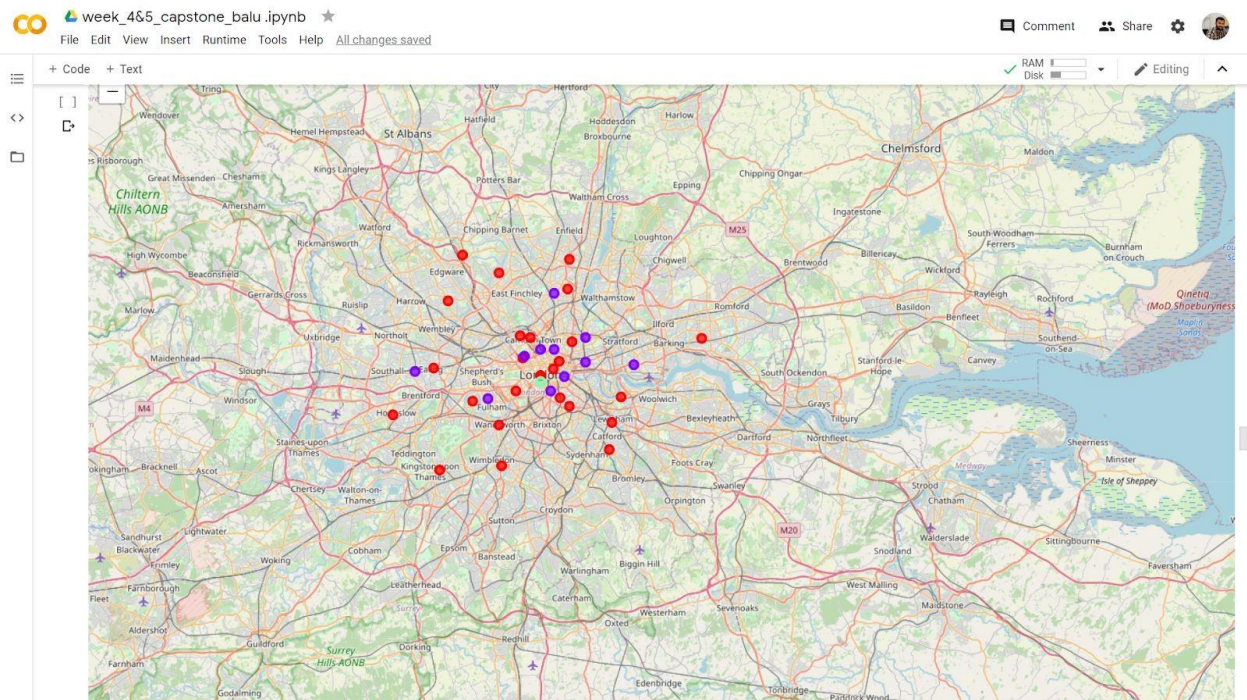
This is the living index cluster map .

This is the cluster for the bars in the city of london.

Discussion :

The living indices of the clusters as known from lon_data_new_mean gives us

0 =7.380769 -**BEST PLACE TO START LIVING IN THE CITY OF LONDON ***

1 =7.315152---**SOME WHAT LESSER SATISFACTION THAN THE CLUSTER 0****

2 =7.300000 ---**BETTER CHOOSE FROM FIRST TWO ****.

These clusters helps us to choose our living neighbourhood for higher living indexThis is very much needed information for almost every businesses that rely on human health and human wellbeing .for eg:if people are healthy and happy ,there is more likely less crime rate so less no police personnel required ,more schools ,more houses in demand in such areas ,more day care centres to name a few .

The results from the k-means clustering of bars show that we can categorize the neighbourhoods into 3

clusters based on the frequency of occurrence for "bars"

• Cluster 0: Neighbourhoods with high number of bars

• Cluster 1: Neighbourhoods with moderate number to no existence of "bars"

• Cluster 2: Neighbourhoods with low concentration of shopping malls

Limitations and Suggestions:

In this project, we only consider one factor i.e. frequency of occurrence of "Bars", there are other factors such as population and income of residents that could influence the location decision of a new shopping mall. However, to the best knowledge of this researcher such data are not available to the neighbourhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new shopping mall. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of a paid account to bypass these limitations and obtain more results.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new Bars. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 1 are the most preferred locations to open a new shopping mall. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open newBars.
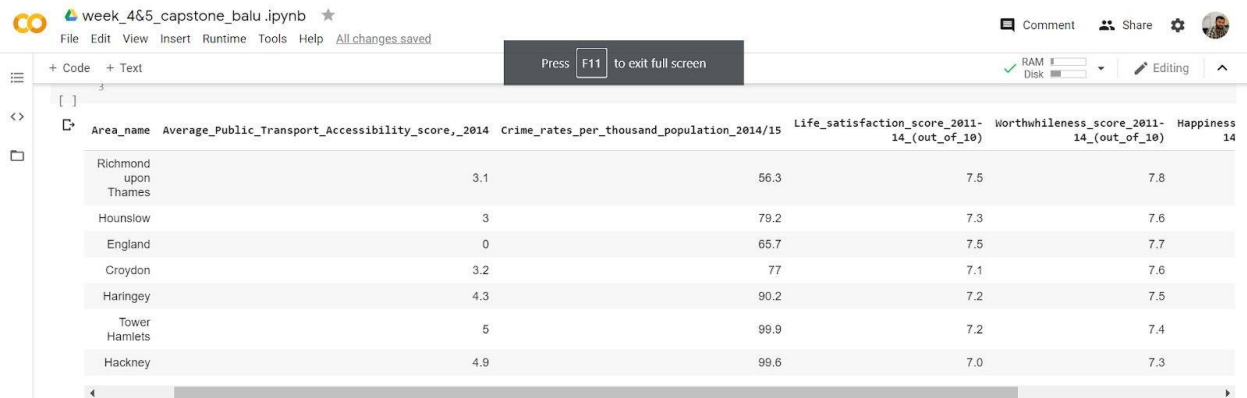
References

London Datastore

:https://data.london.gov.uk/download/london-borough-profiles/c1693b82-68bee-beb2-3decf17dc1f8/london-borough-profiles.csv"

Appendix

Obsevations:All the areas in the city of London ,England have been clustured into three regions as shown above in the map .

They have been clustured based on the *happiness index,worthwhileness index ,lifesatisfaction* indices that give the over all* **living index*** of these areas .

The living indices of the clusturs as known from lon_data_new_mean gives us

0 =7.380769 ->

*BEST PLACE TO START LIVING IN THE CITY OF LONDON *

1 =7.315152--->

CO " week 46,5 capstone balu .ipynb ★

Comment　Share　⚙

RAM
Disk

✎ Editing

+ Code　+ Text

Press F11 to exit full screen

E"	g	Barking and Dagenham	0.00 00 00
		Barnet	0.000000
	3	Bren l	0.011111
	5	Camden	0.000000
	6	City of London	0.020000
	9	Enfield	0.010000
	12	Hackney	0.010000
	13	Hammersmith and Fulham	0.030000
	15	Harrow	0.010000
	18	Hounslow	0.000000
	21	Kensington and Chelsea	0.010000