Coursera : Capstone project

IBM datascience professional certificate

Title: Opening a big chain grocery store in new neighbourhoods

March 2019

Introduction

The Globe is rapidly urbanizing. More than half of the world's population lives in cities, and by 2050 – if current trends continue – the urban population in developing countries is expected to nearly double to 6.7 billion, according to the U.N. Hence it only seems natural that cities are expanding increasingly. And once a deserted and empty area of a city is now rising with population. In my city, Casablanca, not rising areas have a big grocery chain market where they can have legions of choices though they are highly populated. And are left with small retailers that impose super high prices and not many brands to choose from. Consequently, the majority of people living in those areas have to either drive or take public transportation and go to whatever nearest market hence large crowds are shopping at the same time which makes the whole experience excrutiating and unecessarly long. It is evident that opening a new super market is more complicated than one might think as with any business decision. It takes serious work and consideration and one of the most crucial decisions is where to locate the project and where exactly it would be more benificial to build it.

Business problem

The objective of this project is to deploy the tools acquired along the whole course such as machine learning’s techniques to find a solution to a business problem. And the one we will be solving in this project is how to find the best location for super markets in all the rising neighbourhoods if a property developper is looking to open a new supermarket.

Target audience

This project is targeted at any inverstors or property developesr that would like to open a new super market in the capital city of Morocco, Casablanca. This project is timely as the city is currently suffering from oversupply of super markets.

Data

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

→ List of neighbourhoods in Casablanca. This defines the scope of this project which is confined to the city of Casablanca, the capital city of the country of Morocco in North Africa.

→ Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.

→ Venue data, particularly data related to super Markets. We will use this data to perform clustering on the neighbourhoods.

Sources of data and methodolgy to go about this problem

The following wikipedia page https://en.wikipedia.org/wiki/Category:Neighbourhoods_of_Casablanca) contains a list of neighbourhoods in Casablanca, with a total of 24 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages.

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. After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database 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 Super Market 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).