The Battle of Neighbourhoods
(Part-1)

Helping an Expat to find a safer neighbourhood to live in :)

Business Problem

Regardless of where you move, some expat problems always seem to pop up. But knowing how to overcome them will help you embrace your exciting new life.

Moving abroad is a big decision and not one to be taken lightly. To leave your comfort zone and be thrust into an entirely new environment, with its own climate, culture, and customs, is a bold move – to say the least. With all these sudden changes, it is hardly surprising that expat problems arise during the tricky adjustment period. The most important ones are to identify a location to stay which is a safe location, has venues that we are looking for & that fits our budget.

Once the initial culture shock fades away, and you overcome these challenges, expat life can bring numerous rewards; the opportunity to learn a new language, travel the world, and make life-long memories – to name a few. So, to help you make that positive transition, and better prepare yourself to move for the quest that lies ahead, here is a guide to show you some useful insights that can answer the most common expat grievances.

Aim

The aim of the project is to look at the different areas ('towns/townsland') in the Dublin City and classify locations into different clusters based on the demographic, housing-rental index, venues & crime rate information of each location. Thus, helping a person to decide on a cluster to choose.

Introduction

An Expat planning to move to a new country, would ideally have to explore many areas in the city & then decide on a place to live in. The decision is based on various factors of the exploration like:

- Population in the area.
- Crime rate in the area.
- Average Rental Index in the area.
- Popular venues in the area (Parks, Restaurants, Educational Institutions etc).
- Transport options to travel to work.
- places to visit nearby & explore.

This exercise here, tries to achieve the aim of the project using the iterative system of methods that guides on the ideal approach to solving problems with data science. In other words, we can say the project uses the Data Science Methodology, to help solve the problem.

Methodology Used - DATA

Using the Data Science Methodology, I have given the description of the data and how it will be used to solve the problem. ***How the Data is being used at each step is given is highlighted***.

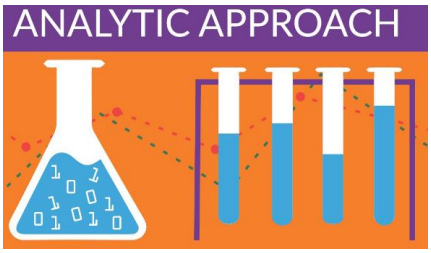
Business /Requirements understanding

What is the problem you are trying to solve?
Here we need to get a clear understanding of what we are trying to achieve in the whole exercise.

Help an Expat to identify the best location to move in.

Analytical approach

How can you use the data to answer the question?



Once a strong understanding of the question is established, the analytic approach can be selected. This means identifying what type of patterns will be needed to address the question most effectively.

Here we are trying to learn about human behaviour, where with the given insights to the data will help them to take a decision. Hence it is appropriate to use Clustering Association approaches.

DATA (Requirements, Collection & Understanding)

What data do you need to answer the question?

Identifying what data is required is very import as the methods of analysis used can require specific content, formats, and data representations.

Where is the data coming from (identify all sources) and how will you get it?

The next important step is to identify the data or Source the data. The data is collected from various data resources (structured, unstructured, and semi-structured) that are relevant to the problem area.

Is the data that you collected representative of the problem to be solved?

In this step Descriptive statistics and visualization techniques can help us understand the content of the data, assess its quality, and obtain initial information about the data.

These 3 steps are iterative and can be ongoing. If any gaps are identified in the data collection/requirements, we may need to repeat the steps and collect more data, fill the gaps as and when required.

For my exercise, data needed are Crime Data in Dublin City, Population Data, Average Rent in each area, Areas or Towns list with Postal Codes in Dublin & location details of each area. The data is collected from various sources like, dataset available in Kaggle for 'Recorded Crime in Ireland from 2003 to 2019 Garda-wise Quarterly', datasets in gov.ie & cso.ie websites along with dataset prepared with the list of areas in Dublin City data from the wiki pages & Foursquare data for getting popular venues.

Data Preparation

Is the data ready to solve the problem or do we need to manipulate the data?

The Data preparation step includes all the activities used to create the dataset used during the modeling phase. This includes cleansing data, combining data from multiple sources, and transforming data into more useful variables.

For this project, the data collected from various data sources is combined and the statistics of each location in Dublin is prepared. The dataset is then cleaned to have the data from 2016 to get better insights for the past 5 years. Risk zone category for each area is given & visualized based on the High, Medium & Low risk areas.

Modeling

In what way can the data be inferred to get the required answer?
The prepared data set is used a train & develop predictive or descriptive models using the described analytical approach previously.

In this project, I have used the K-MEANS Clustering algorithm to determine the clusters of neighbourhoods that are similar in terms or Crime rate, Venues, Population. Along with this KNN classification algorithm is used to train the model to identify the risk areas if a similar dataset is provided with demographic information.

Evaluation

Does the model used really answer the initial question?
Evaluation is a method of checking the quality of the model and helps to verify whether the business problem is handled in a complete and adequate manner or not. This ML technique helps us to identify relationships and trends in data that might otherwise not be accessible or identified.

since we are using k-means, which requires k as an input and does not learn it from data, there is no right answer in terms of the number of clusters that we should have in any problem. We can evaluate how well the models are performing based on different K clusters.

The **Deployment** and **feedback** steps for this project are not implemented as it is a self-study project. You can find the code and this document in the git hub shared in the Coursera for review.