

The Battle of Neighbourhoods

(Part-1 & Part-2)

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/townslans') 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.

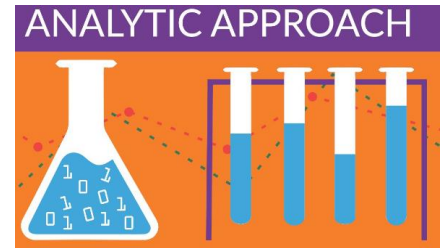
Help an Expat to identify the best location to move in.

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 important 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 to 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 of 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.

Implementation

Data Requirements

- Population data for Dublin City.
- Crime rate data for Dublin City.
- Average Rental Index for Dublin City.
- Popular venues in the area (Foursquare Data).

Data Collection

Dublin areas with estimated population:

This dataset is prepared with the combination of list of areas in Dublin City by postal code and Garda divisions data from the wiki pages & the Census 2016 estimate by townlands available in the cso.ie website.

Dataframe for the same in code looks like:

```
# Above cell is removed when the file is uploaded to github as it contains the credential information.
dublin_areas = pd.read_csv(body)
dublin_areas.head()
```

	Area Name	Postal District	Grada Division	PopulationCensus2016	2021PopulationEstimate
0	Abbey St	Dublin 1	D.M.R. North Central Division	8014	8441
1	Amiens St	Dublin 1	D.M.R. North Central Division	6776	7137
2	Capel St	Dublin 1	D.M.R. North Central Division	1052	1108
3	Dorset St	Dublin 1	D.M.R. North Central Division	1703	1794
4	Henry St	Dublin 1	D.M.R. North Central Division	6300	6636

Crime data of Dublin:

This dataset is downloaded from Kaggle. The dataset in Kaggle is available with the name 'Recorded Crime in Ireland from 2003 to 2019 Garda-wise Quarterly'.

```
crime_data = pd.read_csv(body)
crime_data.head()
```

	REGION	GARDA DIVISION	OFFENCE CODE	OFFENCE	TYPE OF OFFENCE	2003Q1	2003Q2	2003Q3	2003Q4	2004Q1	...	2017Q2	2017Q3	2017Q4	2018Q1	2018Q2	2018Q3	2018Q4	20
0	NORTHERN REGION	CAVAN/MONAGHAN	111	Murder	HOMICIDE OFFENCES	0	0	0	0	0	...	0	0	1	0	0	1	0	
1	NORTHERN REGION	CAVAN/MONAGHAN	112	Manslaughter	HOMICIDE OFFENCES	0	0	0	0	0	...	0	0	0	1	0	0	1	
2	NORTHERN REGION	CAVAN/MONAGHAN	113	Infanticide	HOMICIDE OFFENCES	0	0	0	0	0	...	0	0	0	0	0	0	0	

Average Rental Index data:

This dataset is from & cso.ie website that gives the RTB – average rental index in various locations in Ireland.

```
rental_index_data = pd.read_csv(body)
rental_index_data.head()
```

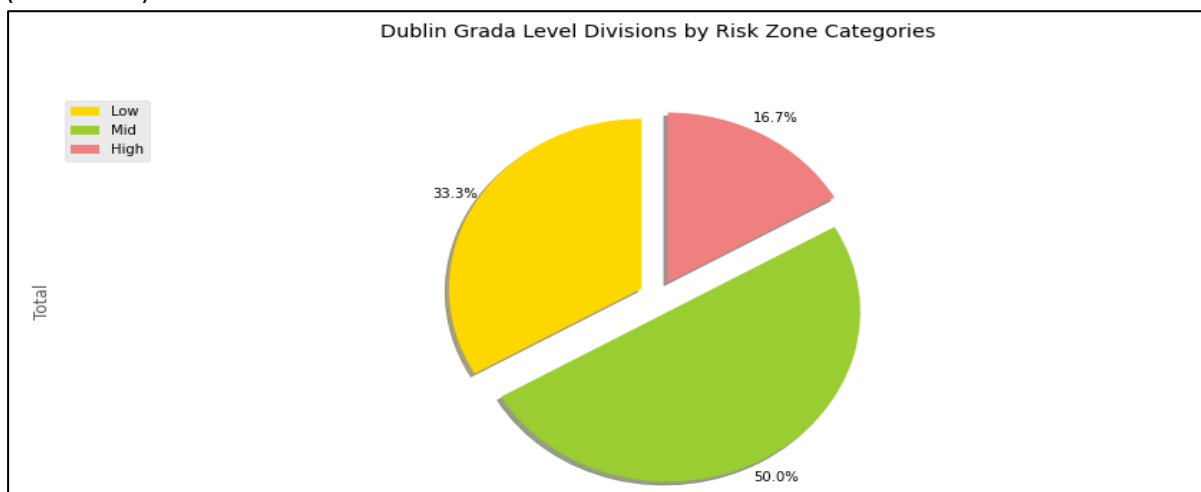
	STATISTIC	Statistic	TLIST(A1)	Year	C02970V03592	Number of Bedrooms	C02969V03591	Property Type	C03004V03625	Location	UNIT	VALUE
0	RIA02	RTB Average Monthly Rent Report	2008	2008	-	All bedrooms	-	All property types	110000	Carlow	Euro	752.82
1	RIA02	RTB Average Monthly Rent Report	2008	2008	-	All bedrooms	-	All property types	110200	Carlow Town	Euro	796.03
2	RIA02	RTB Average Monthly Rent Report	2008	2008	-	All bedrooms	-	All property types	110300	Graiguescullen, Carlow	Euro	754.20

Data Understanding

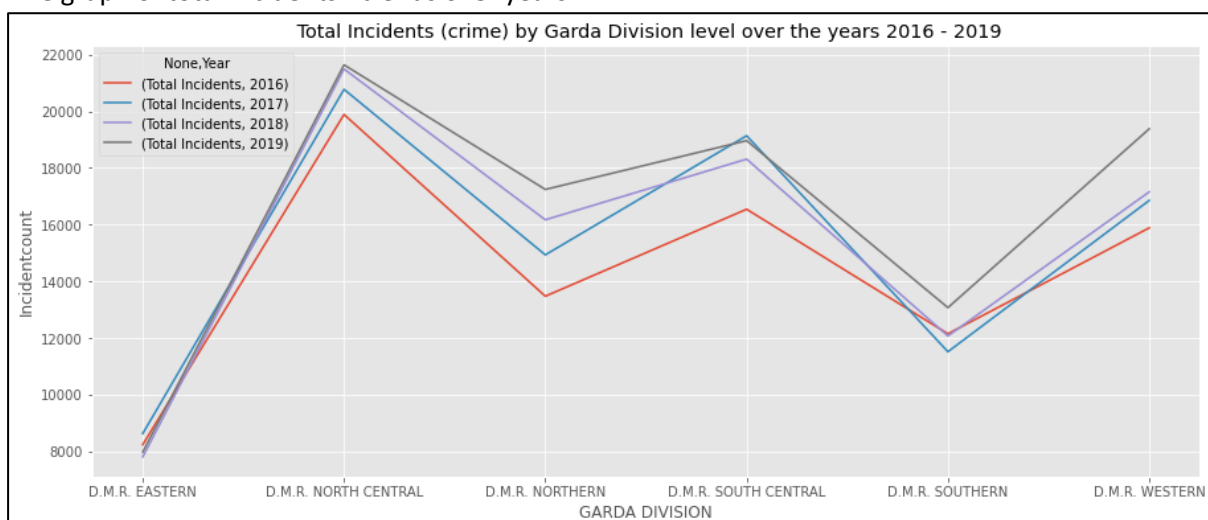
Crime data of Dublin:

Understand the total incidents and based on incidents what risk zone the garda division falls in. This way of visualization helps us to understand the location and its crime rate.

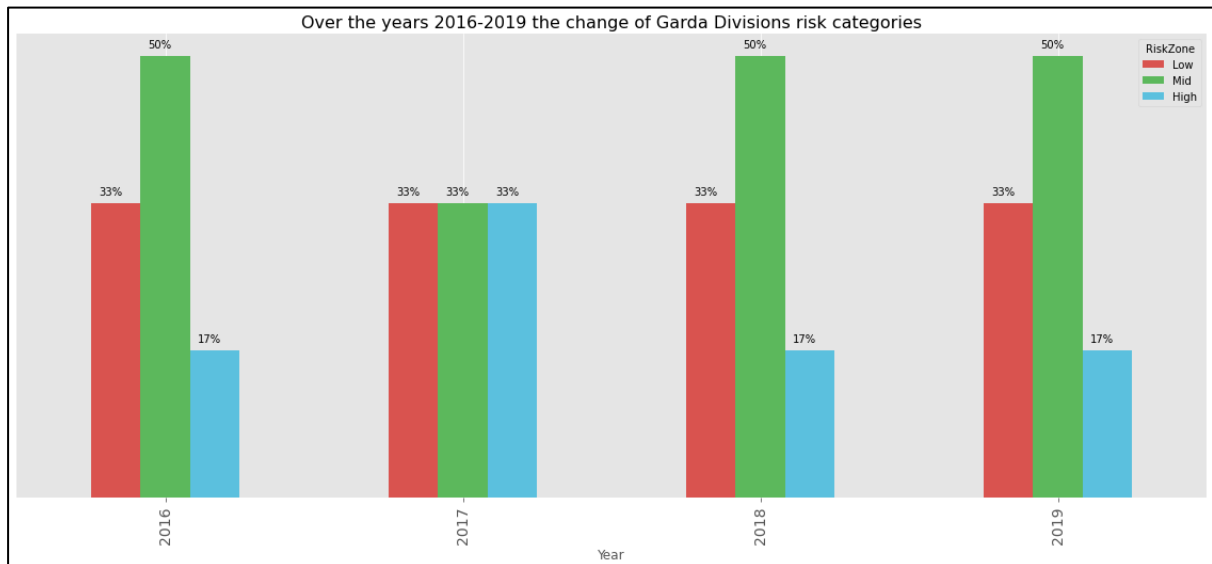
Pie chart for garda division and to which risk zone it falls based on the total incidents in the 4 years (2016-2019)



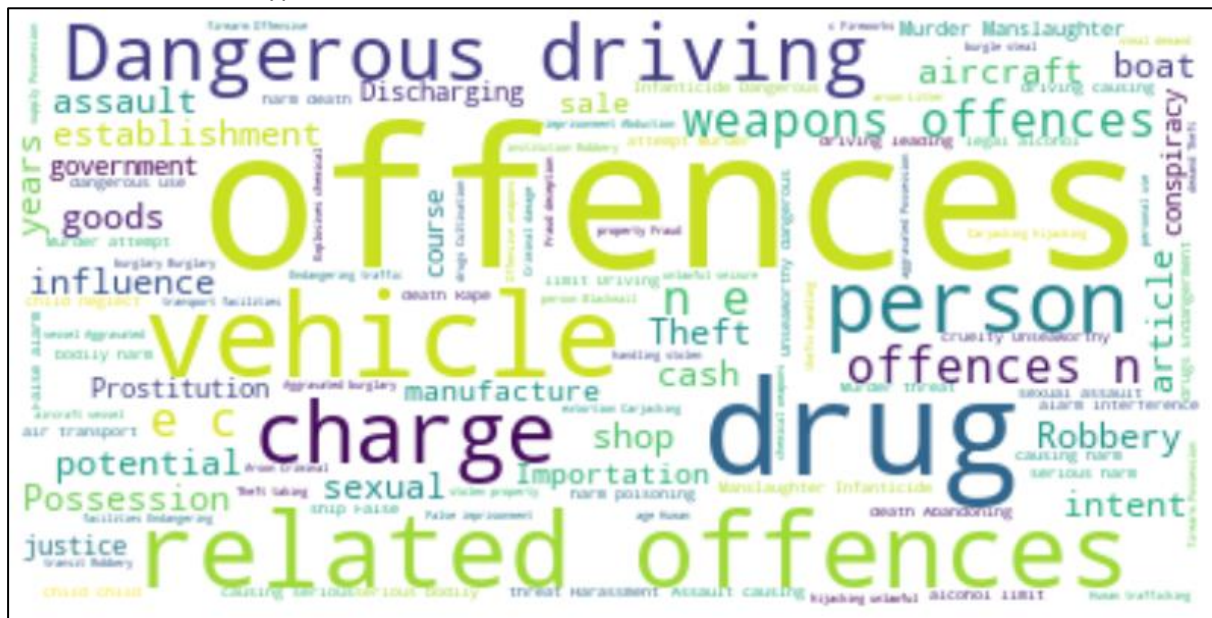
Line graph of total Incidents - trends over years



Bar graph of Risk Zones (count of garda divisions falling in a risk zone) over the years 2016-2019:

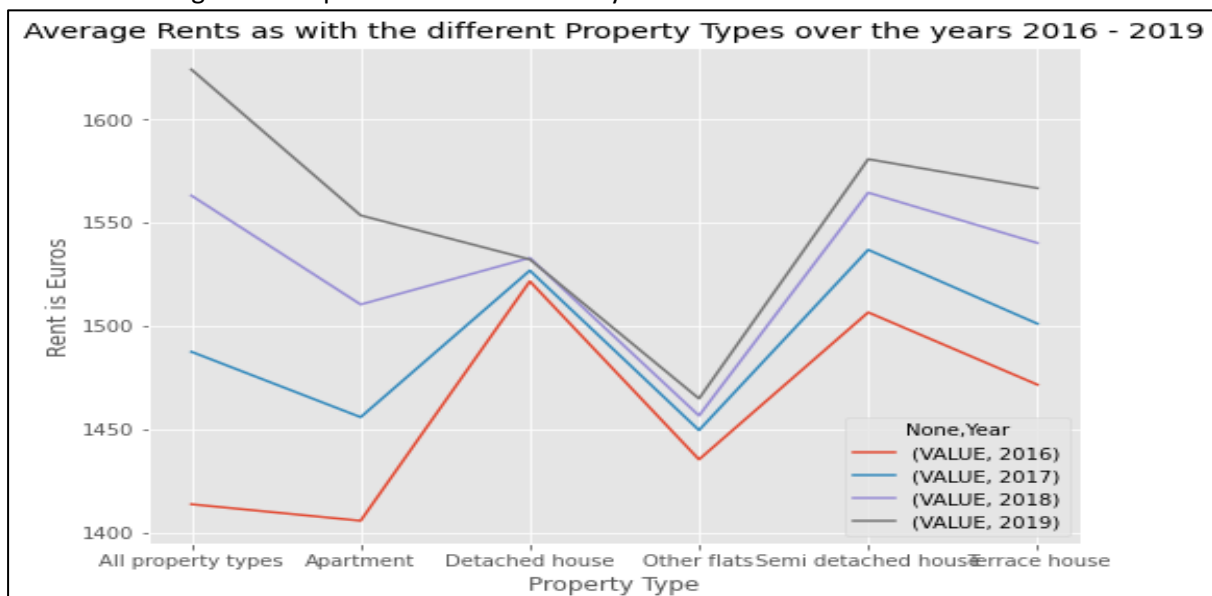


What are the main types of crime incidents that happen in Dublin? This can be understood by a word cloud of all the type of incidents.



Average Rental Index data:

Trends in Average rents in postal districts over the years.



Dublin areas with estimated population Dataset:

Plotting the Locations in a map (Postal Districts of Dublin)



Use of Foursquare data to get the venues in each Postal District of Dublin City

Dublin 8
Dublin 9

12]:

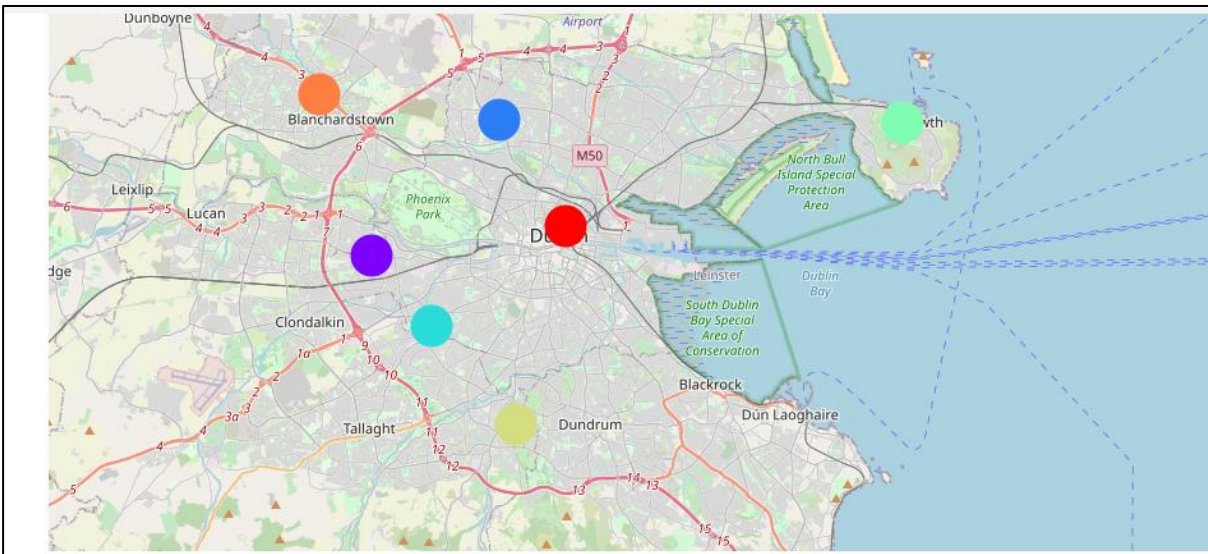
	PostalDistrict	PostalDistrict Latitude	PostalDistrict Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Dublin 1	53.352488	-6.256646	147 Deli	53.353410	-6.259807	Deli / Bodega
1	Dublin 1	53.352488	-6.256646	Holiday Inn Express Dublin City Centre Hotel	53.352306	-6.260955	Hotel
2	Dublin 1	53.352488	-6.256646	The Celt	53.350442	-6.255071	Pub
3	Dublin 1	53.352488	-6.256646	The Hop House (Kim Chi)	53.352981	-6.260772	Korean Restaurant
4	Dublin 1	53.352488	-6.256646	Murray's Bar	53.352419	-6.261256	Pub

Modeling

The original exercise is to find the clusters for so we will use the Clustering Algorithm:

K-Means Clustering

Now that we have all the information, we need to find the clusters based on the data provided. The clusters are formed based on the various datapoints available in each Postal District of Dublin City.



Evaluation & Conclusion

Each cluster gives all the details required for a person to look in. This way we can help solve the problem and can help them Identify a place to move in.

If a person must move in, he/she can examine the clusters formed with the details of the main Venues like Pharmacies, Grocery Stores, Departmental Stores, Parks & many more available. Also, in the Clusters formed, we can get the information on average rents, population details crime incident counts in the latest available year (2019) & based on that what is the Risk Zone.

This way we help an individual to give insights to all the necessary information so that an action can be taken.

The inference made here is clearly based on the dataset available and can change based on the data.

```
: Cluster1 = dublin_merged[dublin_merged['Cluster Labels'] == 0]
Cluster1
```

85]:

	PostalDistrict	GARDA DIVISION	RiskZone	AverageRentalValue	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	...	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
11	Dublin 10	D.M.R. WESTERN	Mid	1351.007000	53.343217	-6.360964	0.0	Bar	Grocery Store	Thai Restaurant	...	Gym	Food Court	Food & Drink Shop	Fish Market	Fish & Chips Shop
29	Dublin 14	D.M.R. EASTERN	Low	1959.605625	53.289133	-6.283512	0.0	Pharmacy	Grocery Store	Fish Market	...	Discount Store	Donut Shop	Electronics Store	English Restaurant	Farmers Market
106	Dublin 6W	D.M.R. SOUTHERN	Low	1666.158824	53.309282	-6.299435	0.0	Pharmacy	Park	Sandwich Place	...	Lake	Bar	Fast Food Restaurant	Fish Market	Fish & Chips Shop