Dublin City Data Analysis

Optimal location for new Grocery Store Business

By Kevin Curtis

Contents

The Business Problem

Many new businesses fail to make profit within their first three years of opening (59% in the hospitality and services industry, according to a study by Dr. HG Parsa). There are many reasons for a business failing to turn a profit. Some of these reasons are out of the control of business owners, while other reasons for failure are entirely under the control of business owners. One area of concern that a business does have a lot of control over is the business location. Because the location of a business in the service industry has such a strong impact on whether that business succeeds or fails, it is an area that demands research and analysis.

A particularly difficult area of business is the restaurant/food and grocery industry. This industry is oversaturated and highly competitive. For this reason, new business owners hoping to open a resaurant or grocery store, in an already saturated market, should aim to find a location for opening their business that gives them the best possible chance for success.

Audience

The intended audience for this project is business people who plan to start a new Grocery store business (specifically looking to locate that business in the Dublin city area).

The research may also be of interest to business owners looking to open one or more restaurants in the Dublin city area.

Data Sources

Foursquare will be used to gather information on venues in the locations that will be researched

Geographical location data will be scraped from Wikipedia.com. The Wikipedia page used to scrape the data for this project is located here: https://en.wikipedia.org/wiki/List_of_Dublin_postal_districts

Mapbox API will be used to get geographical latitude and longituide coordinates

Example JSON data returned from Foursquare:

```
"response": {
  "venues": [
    {
      "id": "5642aef9498e51025cf4a7a5",
      "name": "Mr. Purple",
      "location": {
        "address": "180 Orchard St",
        "crossStreet": "btwn Houston & Stanton St",
        "lat": 40.72173744277209,
        "lng": -73.98800687282996,
        "labeledLatLngs": [
          {
            "label": "display",
            "lat": 40.72173744277209,
            "lng": -73.98800687282996
          }
        ],
        "distance": 8,
        "postalCode": "10002",
        "cc": "US",
        "city": "New York",
        "state": "NY",
        "country": "United States",
        "formattedAddress": [
          "180 Orchard St (btwn Houston & Stanton St)",
          "New York, NY 10002",
          "United States"
        ]
      },
      "categories": [
        {
          "id": "4bf58dd8d48988d1d5941735",
          "name": "Hotel Bar",
          "pluralName": "Hotel Bars",
```


The foursquare API returns a JSON list of venues that includes coordinate locations, category, city, state, country.

Proposed solution

This project will analyse the postal code areas of Dublin, Ireland. Data about the given areas will be collected using the Foursquare API. Geographical information will be retrieved from the web using the python library BeautifulSoup as well as the mapbox API.

The data from Foursquare will be used to cluster the areas into groups based on the number of specific amenities in the local areas. This data will then be used to
make recommendations to business owners about the location of optimal area for setting up a new business.

Foursquare sample response

Bellow is a sample JSON response from the Foursquare API. In this project we will be using the latitude and longitude coordinates of restaurants across Dublin city in order to get an idea of how saurated the grocery store market is within each postal code. We can then use this information to look at areas of lower saturation as possinble locations for a new business.

Methodology

The following libraries were used:

- BeautifulSoup
- Requests
- Matplotlib
- Pandas
- Folium

Geographic data was scraped from wikipedia, using BeautifulSoup to get all of the Dublin city area codes along with a list of place names within each area code.

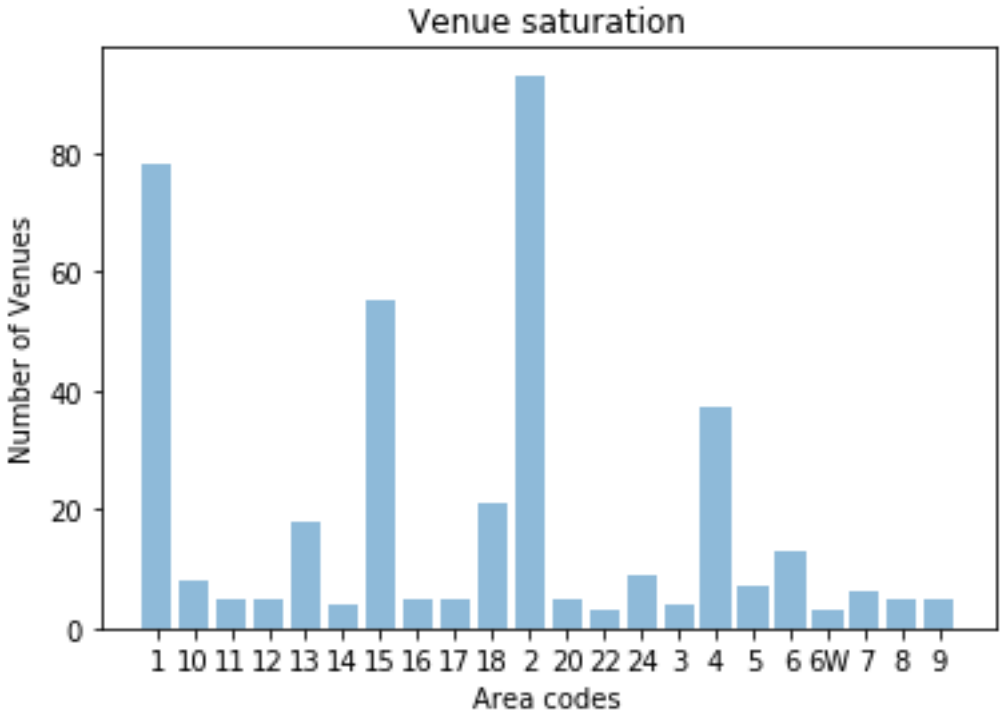
The Geographic data was made into a dataframe, using pandas, in order to allow for easy manipulation and better comprehension of the data. See a snippet of the resulting dataframe below:

	Area codes	Place names
0	1	[Abbey Street, Amiens Street, Dorset Street, H...
1	2	[River Liffey, Merrion Square, Trinity College...
2	3	[Ballybough, North Strand, Clonliffe, Clontarf...
3	4	[Ballsbridge, Belfield, Donnybrook, Irishtown,...
4	5	[Artane, Coolock, Harmonstown, Kilbarrack, Rah...
5	6	[Milltown, Ranelagh, Terenure, Rathmines, Dart...
6	6W	[Harold's Cross, Templeogue, Kimmage, Terenure]
7	7	[Arbour Hill, Ashtown, Broadstone, Cabra, Gran...
8	8	[Dolphin's Barn, Inchicore, Islandbridge, Kilm...
9	9	[Ballymun, Beaumont, Donnycarney, Drumcondra, ...
10	10	[Ballyfermot, Cherry Orchard]
11	11	[Ballymun, Finglas, Ballygall, Glasnevin, Glas...
12	12	[Bluebell, Crumlin, Drimnagh, Greenhills, Perr...
13	13	[Baldoyle, Bayside, Donaghmede, Sutton, Howth,...
14	14	[Churchtown, Clonskeagh, Dundrum, Goatstown, R...
15	15	[Ashtown, Blanchardstown, Castleknock, Coolmin...
16	16	[Ballinteer, Ballyboden, Dundrum, Kilmashogue,...
17	17	[Balgriffin, Coolock, Belcamp, Darndale]

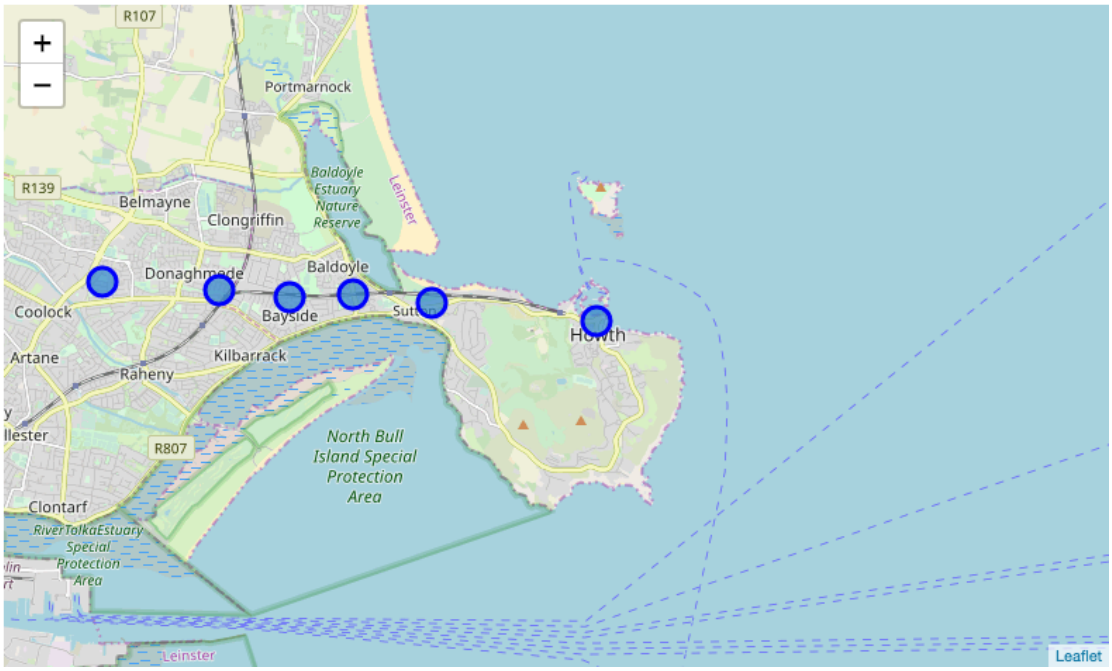
The Mapbox API was used to generate accurate latitude and longitude variables for each area code, using geocoding. These coordinate data were then built into a dataframe, using pandas. The resulting dataframe was then merged with the existing dataframe in order to provide an easily understandable representation of each location and its corresponding coordinates.

	Area codes	Place names	Latitude	Longitude
0	1	[Abbey Street, Amiens Street, Dorset Street, H...	53.352565	-6.256647
1	2	[River Liffey, Merrion Square, Trinity College...	53.338923	-6.252702
2	3	[Ballybough, North Strand, Clonliffe, Clontarf...	53.361246	-6.185515
3	4	[Ballsbridge, Belfield, Donnybrook, Irishtown,...	53.327655	-6.227492
4	5	[Artane, Coolock, Harmonstown, Kilbarrack, Rah...	53.383656	-6.181606
5	6	[Milltown, Ranelagh, Terenure, Rathmines, Dart...	53.317884	-6.259986
6	6W	[Harold's Cross, Templeogue, Kimmage, Terenure]	53.306082	-6.300782
7	7	[Arbour Hill, Ashtown, Broadstone, Cabra, Gran...	53.360430	-6.284417
8	8	[Dolphin's Barn, Inchicore, Islandbridge, Kilm...	53.350404	-6.320374
9	9	[Ballymun, Beaumont, Donnycarney, Drumcondra, ...	53.385823	-6.245707
10	10	[Ballyfermot, Cherry Orchard]	53.343185	-6.361034
11	11	[Ballymun, Finglas, Ballygall, Glasnevin, Glas...	53.386611	-6.292611
12	12	[Bluebell, Crumlin, Drimnagh, Greenhills, Perr...	53.320503	-6.326052

Using the location data gathered, the Foursquare API was queried, which returned a JSON formatted list of venues in each location. The data gathered from Foursquare was visualised as a bar chart, using matplotlib. This gave us an overview of each area code and allowed us to decide on refining our search to areas localised in the Dublin 13 area.



Folium was used to visualise the dublin 13 areas on a map:



The above process was repeated on the smaller data set generated by refining our search to Dublin 13. This dataset was then analysed with the k-means algorithm. This allowed for a decision to be made on the best location for a new grocery store.

Results

The results of our k-means clustering showed that the first cluster (cluster 0) had the lowest percentage of Grocery stores in the clustered areas. The second cluster (cluster 1) had the highest concentration of Grocery stores, and the third cluster (cluster 3) had a low to moderate number of Grocery stores. Regarding these measurements it would be advantageous to open a new Grocery store in either of the areas (Bayside, Donaghmede) that belong to cluster 0.

	Area name	Grocery Store	Cluster Labels
0	Ayrfield	0.095238	2
1	Baldoyle	0.000000	0
2	Bayside	0.142857	1
3	Donaghmede	0.142857	1
4	Howth	0.019231	0
5	Sutton	0.000000	0

Conclusion and Further research

While this project looked at potential areas for opening a new grocery store based on preexisting grocery stores there are many other data sets that should be investigated in order to create a more thorough report. Some of these data may include demographics, affluence, population density, spending trends, and crime rates to name a few.

This project will continue to grow as it integrates more of the above mentioned data sets.