The Missing Irish Pub

IBM Data Science Professional Certificate Capstone Project

Christian Weymann

1 Introduction

In a large city like New York, each neighborhood can have a different feel to it: some are purely residential, others have a lot of shopping opportunities, or business centers, or nightlife venues. Implanting a new business in the wrong type of neighborhood can be a fast road to failure, as customers will not think of going to that location for this type of service. It also drastically reduces the chances of being discovered by accident by a customer who was already in the area.

In this project, we will take the example of Irish pubs to demonstrate this principle. Using data from the Foursquare Places API, we will determine what kind of neighborhood would be a good location for a new Irish pub in New York City. To this end, we will try to predict the number of Irish pubs in each neighborhood based on the frequency of other types of venues present in that area. We can then compare the predicted number to the actual number to identify opportunities.

The same kind of analysis done here for Irish pubs could be done for any type of business. Any prospective business owner could benefit from reproducing it. It could also be used by local governments to try to get a more data backed understanding of the area under their responsibility, and what kind of businesses they need to attract if they want to change the feel of their neighborhood.

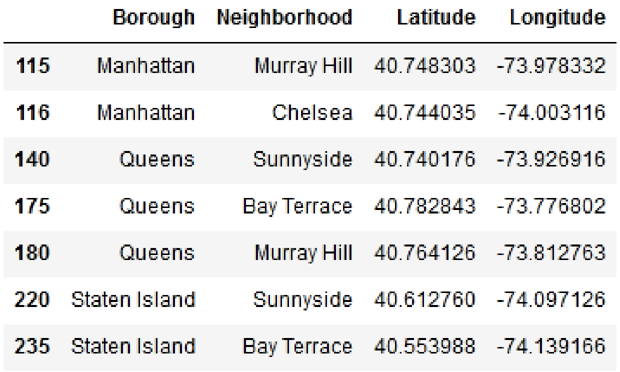
2 Data

2.1 Data Acquisition

Our data will mostly be loaded from the Foursquare Places API. Foursquare maintains a list of venues with a lot of rich data associated with each of them, such as ratings and comments from its users. However, we will be mostly focuse on core data about each venue: its geographical position, and its category. Foursquare lists the category of each venue both with a name and a unique category identifier which can be used in searches. We will take advantage of this to find all the Irish pubs in each neighborhood.

2.1.1 Neighborhood data

The geographical data of each neighborhood is taken from the json file used during the lab session of this course. It contains geographical information for the 306 neighborhoods of New York City. We extract the names of the neighborhoods and their boroughs, as well as the central latitude and longitude. These will serve as the positions around which to search for venues in our calls to the Foursquare API.



	Borough	Neighborhood	Latitude	Longitude
115	Manhattan	Murray Hill	40.748303	-73.978332
116	Manhattan	Chelsea	40.744035	-74.003116
140	Queens	Sunnyside	40.740176	-73.926916
175	Queens	Bay Terrace	40.782843	-73.776802
180	Queens	Murray Hill	40.764126	-73.812763
220	Staten Island	Sunnyside	40.612760	-74.097126
235	Staten Island	Bay Terrace	40.553988	-74.139166

Table 1: List of neighborhoods with the same name

For further analysis, we need a unique way to identify each of the neighborhoods. Using their names is not practical, as there are several pairs of neighborhoods with the same name in different boroughs, see Table 1. Instead, the index of this data frame will serve as a unique identifier for each neighborhood.

2.1.2 Pub data

We acquire the data for the pubs in each neighborhood using the "search" endpoint of the "venues" group of the Foursquare Places API. This endpoint allows us to pass a category ID as an argument to limit the results to venues of a particular category. We use this option to find all Irish pubs within 500 meters of the center of each neighborhood, and load them into a data frame for further analysis.

2.1.3 Typical venues data

To find the typical venues in each neighborhood, we use the "explore" endpoint of the "venues" group. This returns up to one hundred venues that are close to the center of each neighborhood, but are also considered relevant by Foursquare. Furthermore, this endpoint does not require to be passed an explicit radius to be searched, instead defaulting to a suggested radius based on the density of venues in the area. This allows us to get a better picture of what venues are actually typical in each neighborhood.

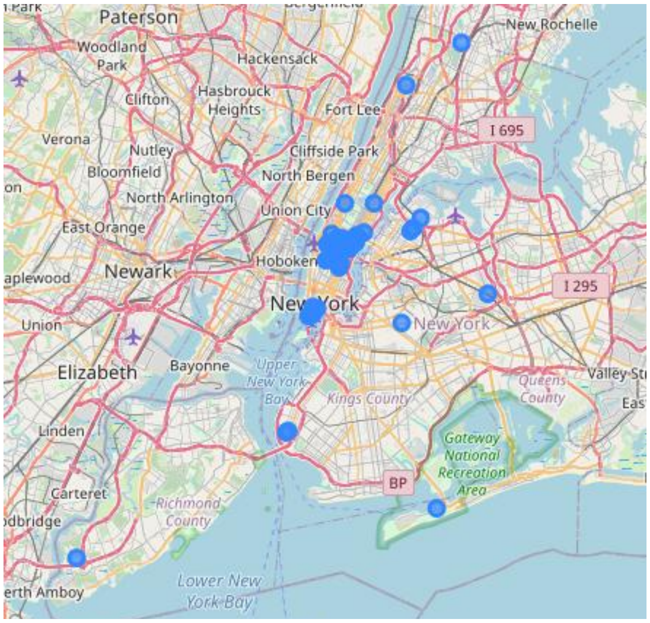


Figure 1: Locations of all the Irish pubs in New York City

2.2 Data Exploration

To get a first idea of our pub data, we plot the position of each pub on a map of New York City (Figure 1). We can see that the pubs are very inhomogeneously distributed, with most of them being located in central Manhattan.

Our target variable is going to be the number of Irish pubs in a given neighborhood. We can easily compute it by counting the entries in our pub data frame that are located in each neighborhood. The distribution of this target variable is shown in Figure 2. As we could already guess from the map, most neighborhoods do not have any Irish pubs, whereas a few neighborhoods have many. This is expected for a variable describing the number of occurrences of a rare event, which is best described using a Poisson distribution.

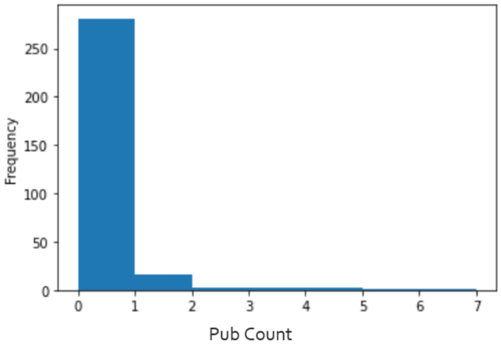


Figure 2: Histogram of the number of Irish Pubs in each neighborhood

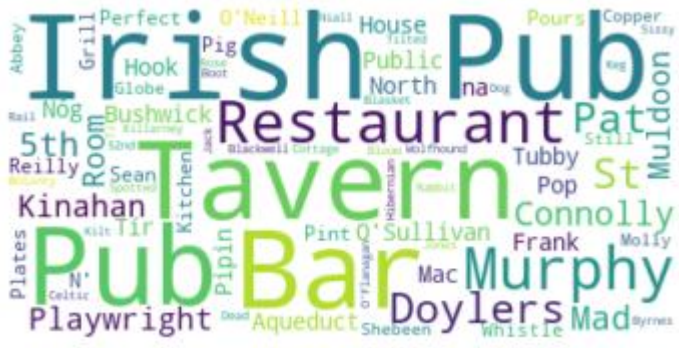


Figure 3: Word cloud generated from the names of the Irish pubs in New York City

Finally, let us search for inspiration to name our new pub. We generate a word cloud from the names of all the pubs in our data frame, shown on Figure 3. Unsurprisingly, the words “Pub”, “Bar”, and “Irish Pub” are very common. “Tavern” seems to also be frequently used, indicating that many of these venues go for an “old timey” feel. Other notable categories are Irish sounding names, and beer brand names. The word “pint” also makes an appearance.