THE MISSING IRISH PUB

IBM Data Science Professional Certificate Capstone Project

Christian Weymann

SO YOU WANT TO OPEN AN IRISH PUB IN NEW YORK CITY?

- We will find a suitable neighborhood for the next Irish pub in New York City by comparing neighborhoods that do and that don't have an Irish pub, and finding one that doesn't, but should, have one

- For each neighborhood, we will search for Irish pubs using the «venues/search» endpoint of the Foursquare API

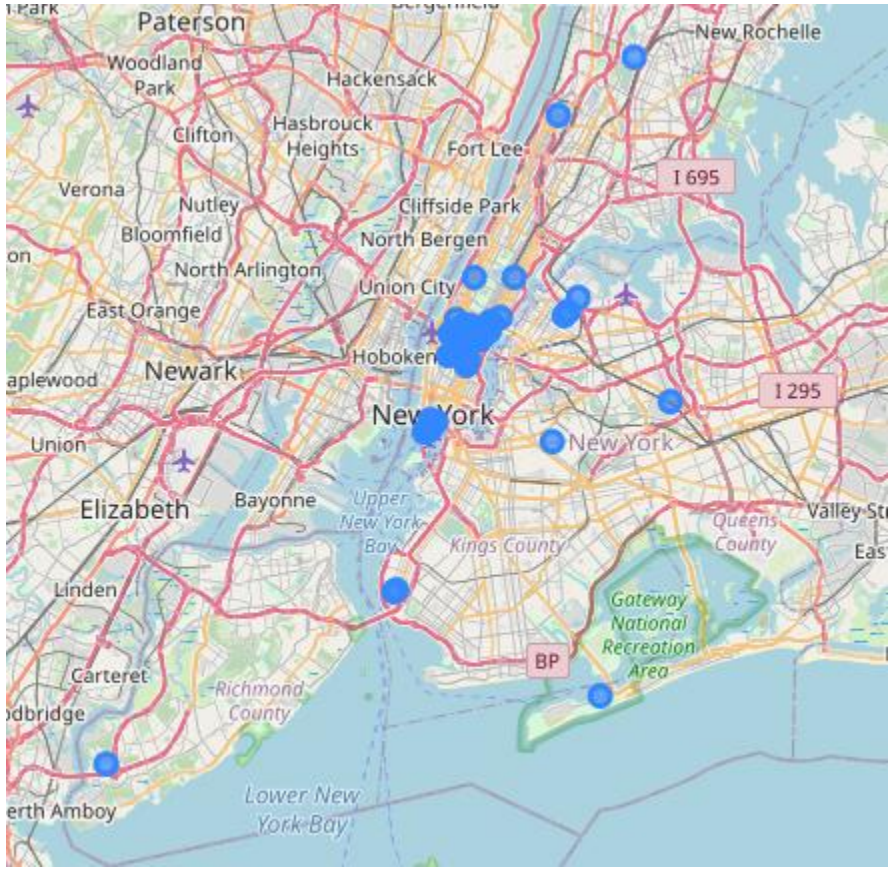
- We will collect typical venues for each neighborhood using the «venues/explore» endpoint of the Foursquare API

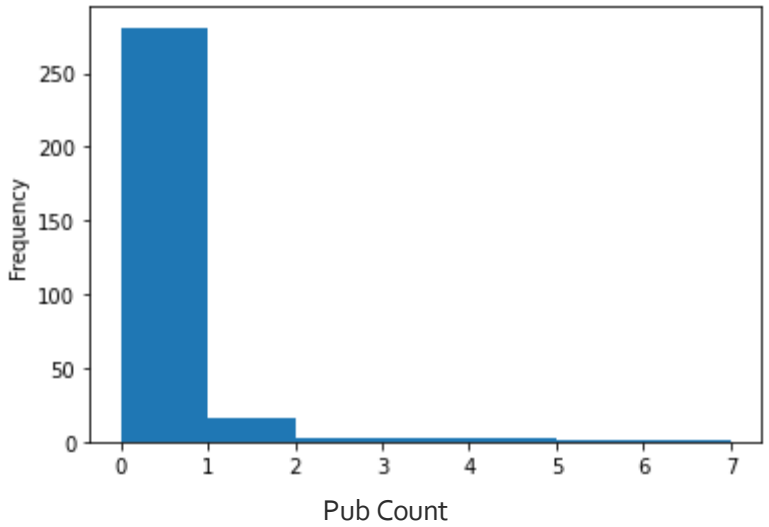
FIRST CHALLENGE: NEIGHBORHOOD NAMES ARE NOT UNIQUE

	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
244	Staten Island	Chelsea	40.594726	-74.189560

- Several pairs of neighborhoods have the same name
- Use the index of a reference data frame as unique identifier

SECOND CHALLENGE: VERY INHOMOGENEOUS DISTRIBUTION OF PUBS





- The map and the histogram show that most neighborhoods do not have an Irish pub
- Distribution will be best described by a Poisson distribution

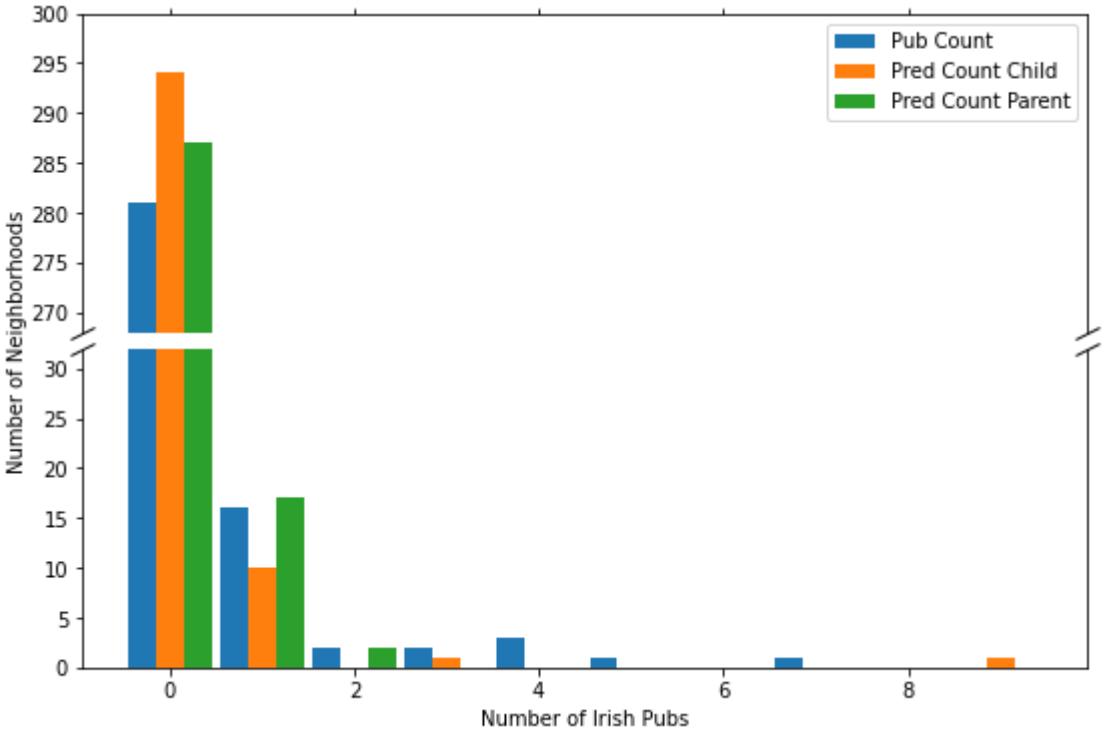
THIRD CHALLENGE: MANY VERY DETAILED CATEGORIES

Parent category	Child categories
Bar	Pub, Bar, Sports Bar, Beer Bar, Wine Bar, Beer Garden, Cocktail Bar, Speakeasy, Karaoke Bar, Hookah Bar, Whisky Bar, Tiki Bar, Dive Bar, Gay Bar, Hotel Bar, Sake Bar, Beach Bar
Movie Theater	Multiplex, Indie Movie Theater, Movie Theater
Stadium	Tennis Stadium, Baseball Stadium, Track Stadium, Basketball Stadium, Stadium

- Many categories are very detailed, and therefore appear very sparsely in our frequency data
- We use the category hierarchy from Foursquare to generate two alternative datasets:
  - «Parent» with grouped categories
  - «Child» with the original categories

BOTH MODELS SEEM TO WORK WELL ON DIFFERENT PARTS OF THE DISTRIBUTION

- The model trained on the «Parent» data seems to reproduce the low pub counts well
- The model trained on the «Child» data fits outliers with a high number of pubs
- However, it overestimates the number of pubs in those outliers, leading it to always recommend these neighborhoods
- Use the «Parent» model for recommendations



FINAL VERDICT: TRIBECA SHOULD GET AN IRISH PUB

- Map shows the top ten recommendations of the «parent» model
- Orange is the highest score, dark blue the lowest
- Tribeca in Manhattan is the neighborhood with the highest recommendation score