The Battle of Neighborhoods

Coursera's "Applied Data Science Capstone" Project

Student: Michael Onishi
Date: June 2020

The Goal

Recommend a list of neighborhoods in San Diego, California, to someone based on his/her interests in the following categories:

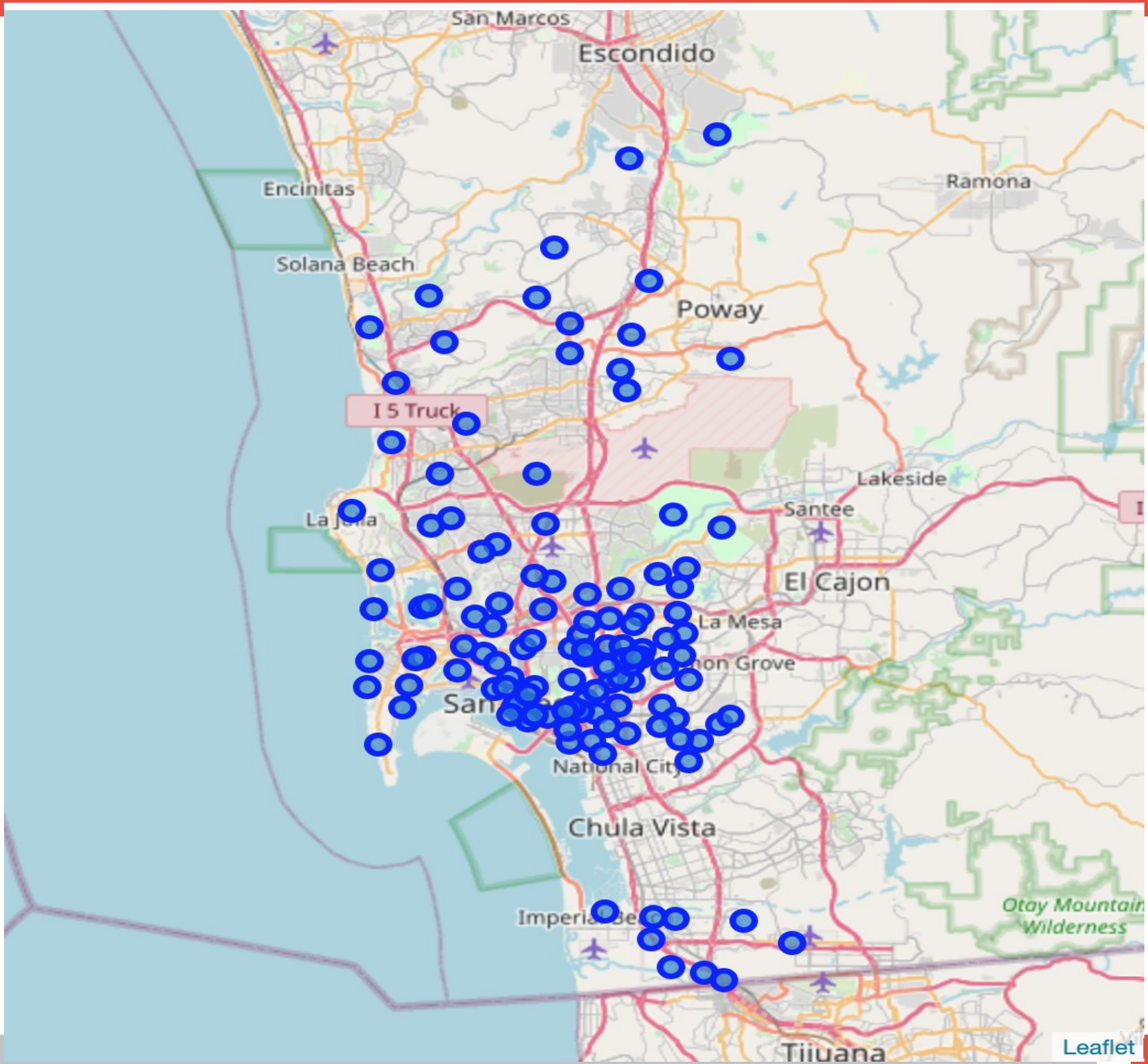
- Arts & Entertainment
- College & University
- Food
- Nightlife Spot
- Outdoors & Recreation
- Shop & Service

Potential users: anyone looking for a place to live in San Diego or just going there for tourism

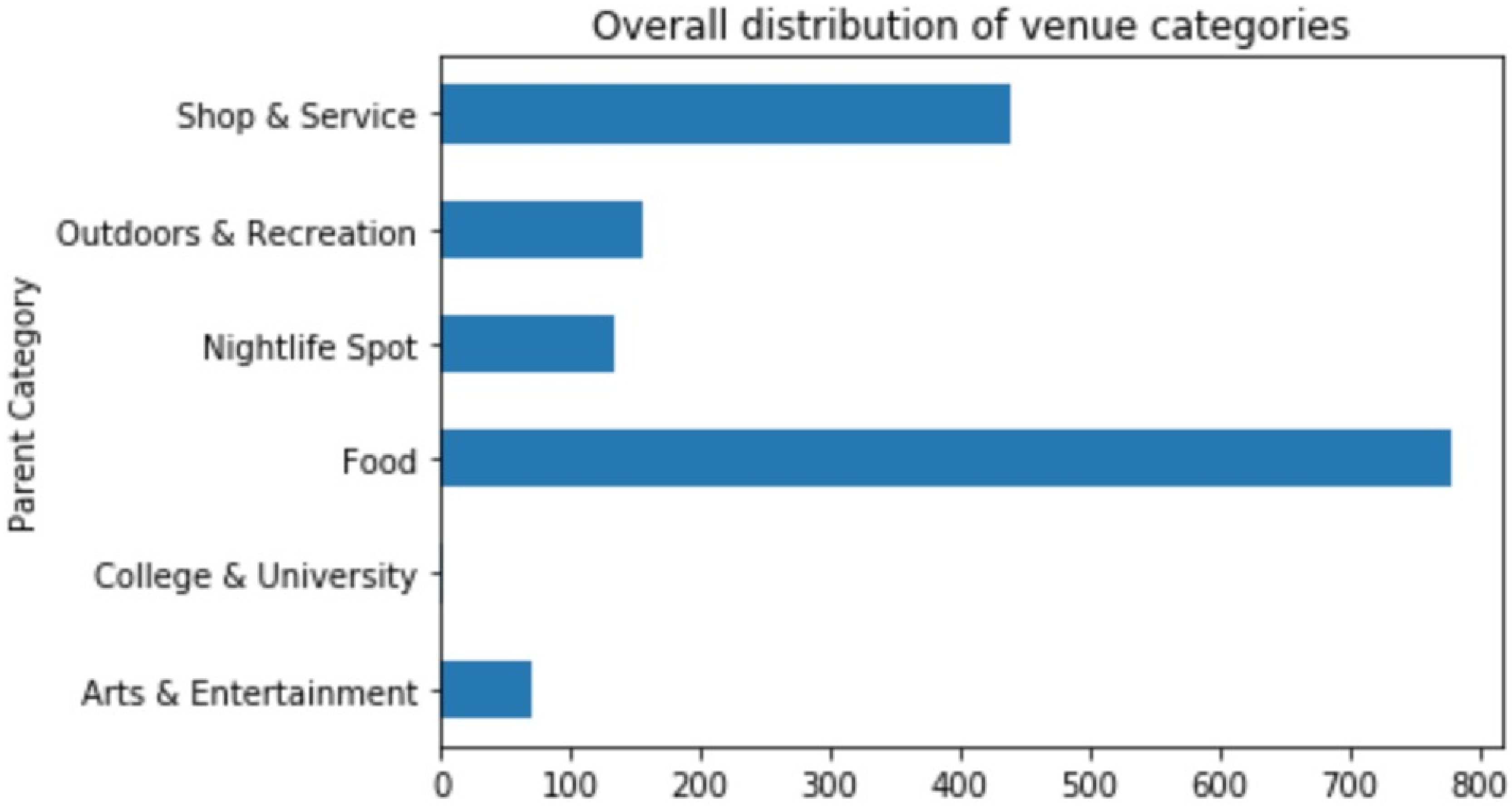
Data aquisition and wrangling

- Neighborhood data from San Diego Open Data Portal (https://data.sandiego.gov/datasets/pd-neighborhoods/)
- Total of 124 neighborhoods, location data in geojson format
- Took the mean of latitudes and longitudes to get one coordinate for each neighborhood
- Got up to 100 venues for each neighborhood
- Got top level venue category for each detailed category
- Aggregated top level venue categories by neighborhood

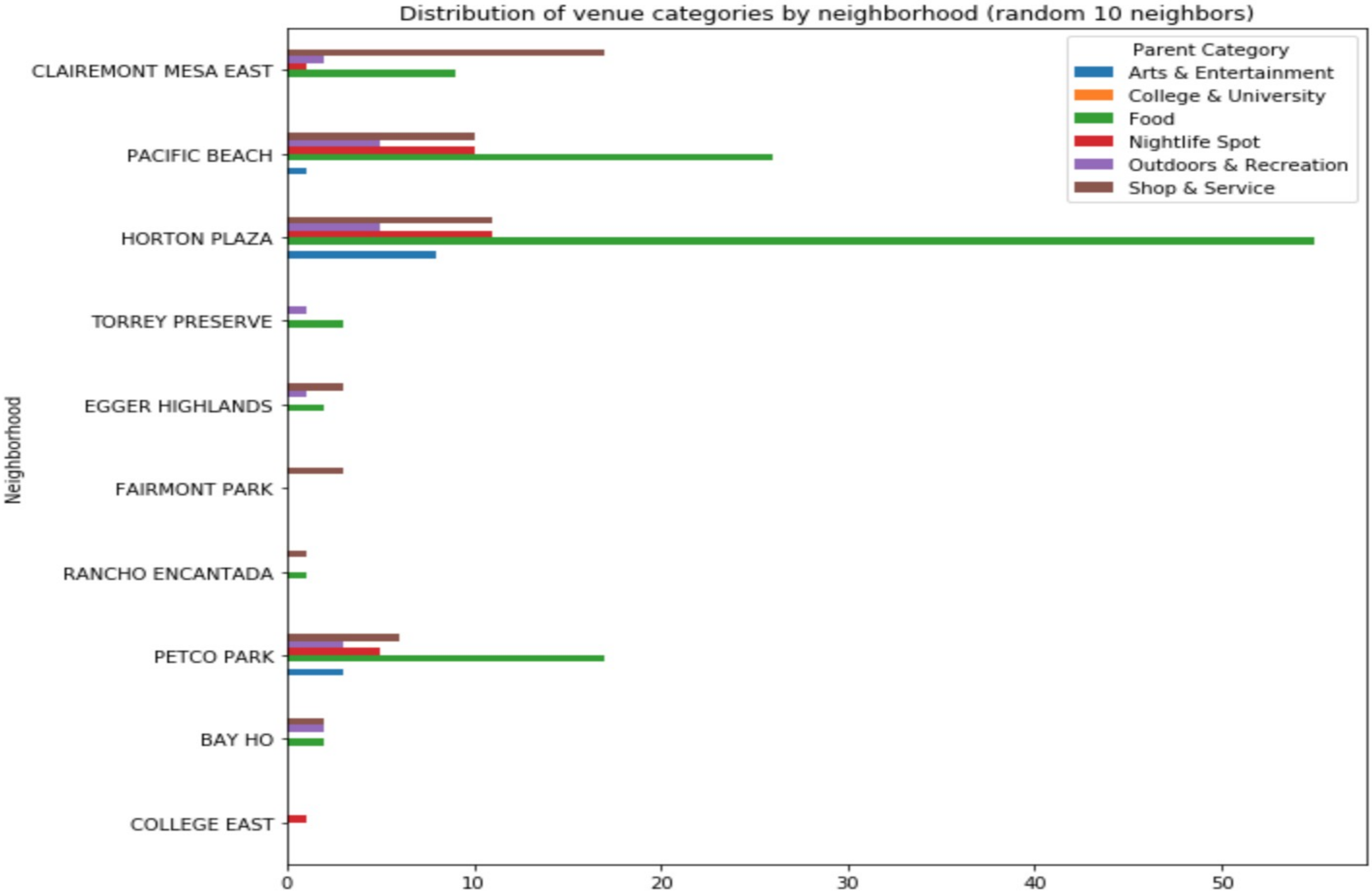
San Diego neighborhoods



Overall distribution of the venue categories



Distribution of categories by 10 random neighborhoods



Created profiles to test

Profile 1: Young single men looking for fun

- Arts & Entertainment = 6
- College & University = 4
- Food = 2
- Nightlife Spot = 10
- Outdoors & Recreation = 5
- Shop & Service = 2

Created profiles to test

Profile 2: A stable mid-aged couple looking for living in the city

- Arts & Entertainment = 7
- College & University = 0
- Food = 6
- Nightlife Spot = 2
- Outdoors & Recreation = 5
- Shop & Service = 6

Created profiles to test

Profile 3: An retired woman looking for peace

- Arts & Entertainment = 7
- College & University = -1
- Food = 5
- Nightlife Spot = -5
- Outdoors & Recreation = 10
- Shop & Service = 5

How to recommend?

Cosine similarity to make sure the disparity on the number of venues do not interfered in the score.

It also maked it possible to the user to choose any range of weights for the categories.

Example:

BALBOA PARK = [9, 0, 4, 0, 9, 0]

Profile 1 = [6, 4, 2, 10, 5, 2]

Cosine similarity = (9*6 + 4*2 + 9*5) / (sqrt(9^2 + 4^2 + 9^2) * sqrt(6^2 + 4^2 + 2^2 + 10^2 + 5^2 + 2^2)) = 0.58964132349

Top 5 recommendations

Profile 1:

Neighborhood	
SERRA MESA	0.778078
ALLIED GARDENS	0.735215
COLLEGE EAST	0.735215
UNIVERSITY CITY	0.666924
SOUTH PARK	0.660330

Top 5 recommendations

Profile 2:

Neighborhood	
ADAMS NORTH	0.979796
LINDA VISTA	0.895669
SAN YSIDRO	0.835573
QUALCOMM	0.833333
CORTEZ	0.818165

Top 5 recommendations

Profile 3:

Neighborhood	
ADAMS NORTH	0.900000
SAN CARLOS	0.870930
BALBOA PARK	0.864460
LOGAN HEIGHTS	0.812362
OCEAN CREST	0.801388

Conclusion

- The recommendations made sense with the available data
- San Diego is not so well covered by Foursquare
- Working with larger areas than neighborhoods may solve the small data size problem
- If a large venue dataset is available, it should be better to collect users preferences on finer grained categories to match better their profile and give better recommendations