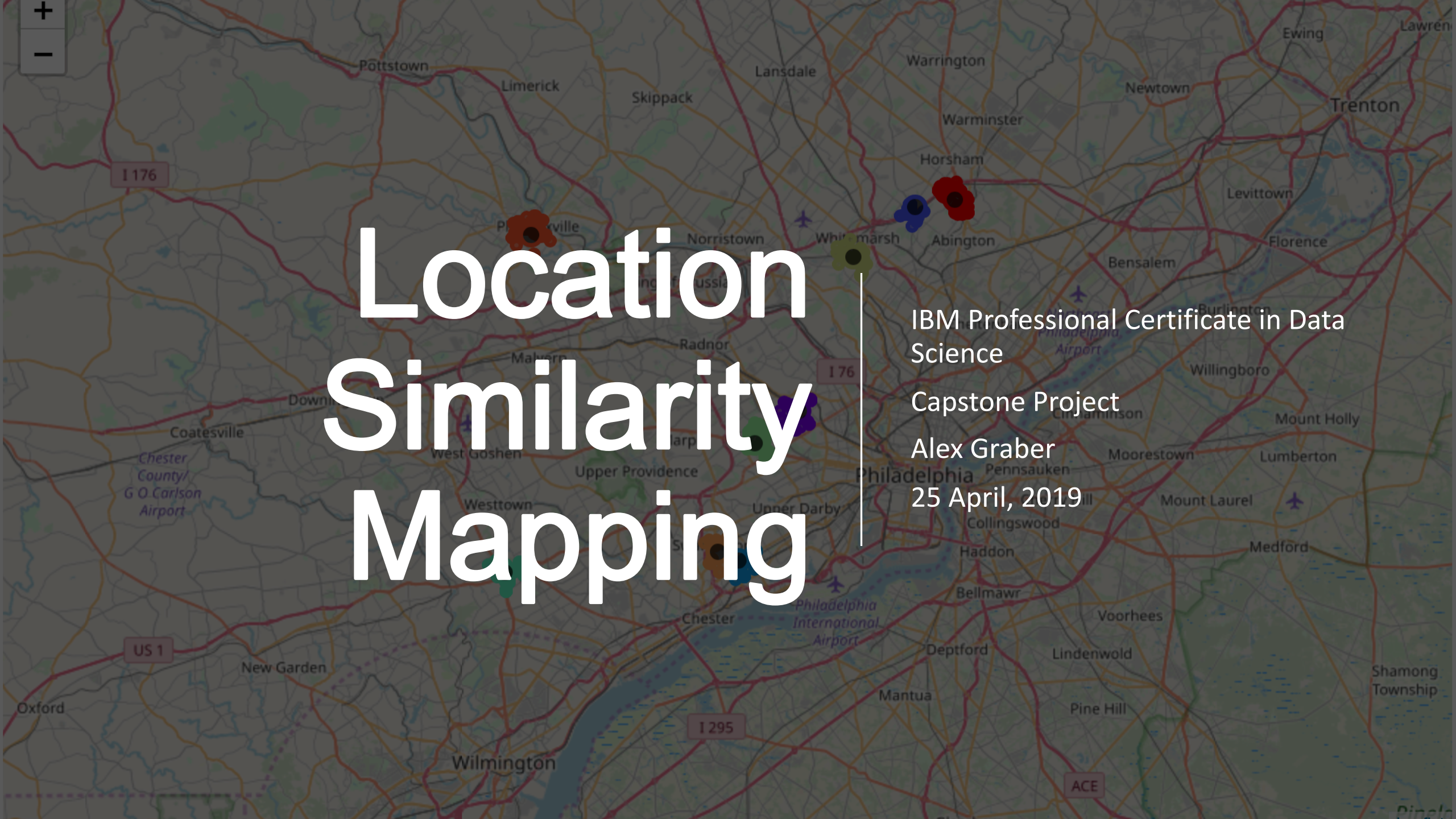

Location Similarity Mapping
IBM Professional Certificate in Data Science
Capstone Project
Alex Graber
25 April, 2019
Pottstown
Limerick
Skippack
Lansdale
Warrington
Newtown
Trenton
Warminster
Horsham
Levittown
Norristown
Abington
Bensalem
Florence
Radnor
Malvern
Willingboro
Coatesville
Chester County/ G O Carlson Airport
West Goshen
Upper Providence
Philadelphia
Pennsauken
Moorestown
Mount Holly
Lumberton
Westtown
Upper Darby
Mount Laurel
Collingswood
Haddon
Medford
Bellmawr
Chester
Philadelphia International Airport
Voorhees
Deptford
Lindenwold
New Garden
Oxford
Mantua
Pine Hill
Shamong Township
Wilmington
I 176
US 1
I 295
ACE

Introduction

If I like where I'm living now, but need to move, in which areas should I begin my search?

In this project, we identify characteristics that define a source zip code (current or idealized) and find the most similar zip codes in a broader destination location.

Data collected from FourSquare and Zillow

In order to understand the similarity between areas, we required several distinct types of data:

- A way to define an area's location geographically.
- A way to define the characteristics of a location.
- A 'source' location used as an example to search against.
- A pool of 'destination' locations used to find the final subset of top destination areas.

Zip code geolocation coordinates were provided via Zillow's 'GetRegionChildren' API. Assuming we know the county (or counties) in which we're interested in living, the Zillow API allows us to collect all zip codes from within these counties.

Once we have zip codes and latitude/longitude coordinates, the FourSquare API allows us to understand the venue characteristics that identify a particular location

Methodology: Data Setup

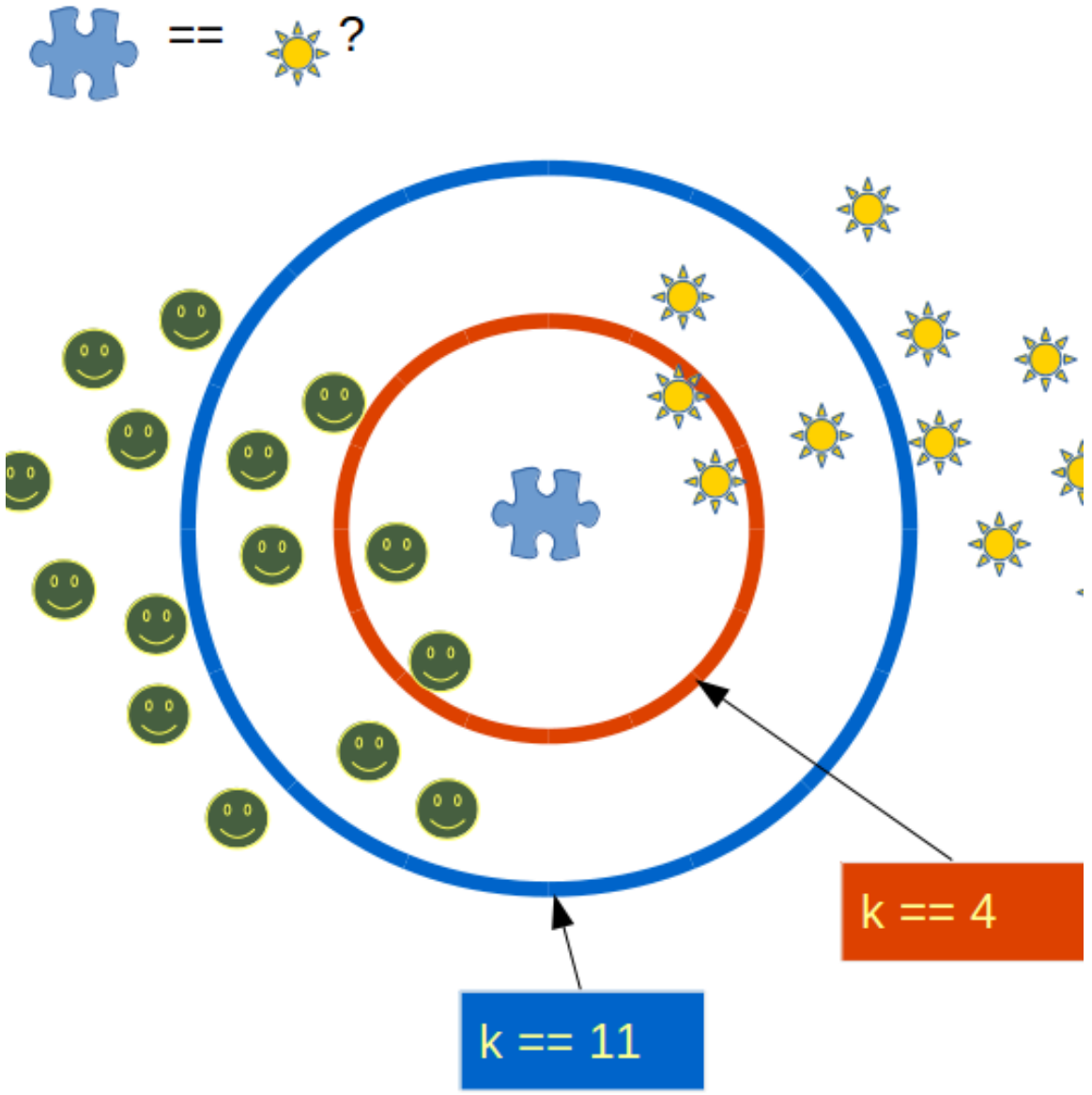
A region is defined by the types of venues within 1600m (~1 mi) of the center of the zip code

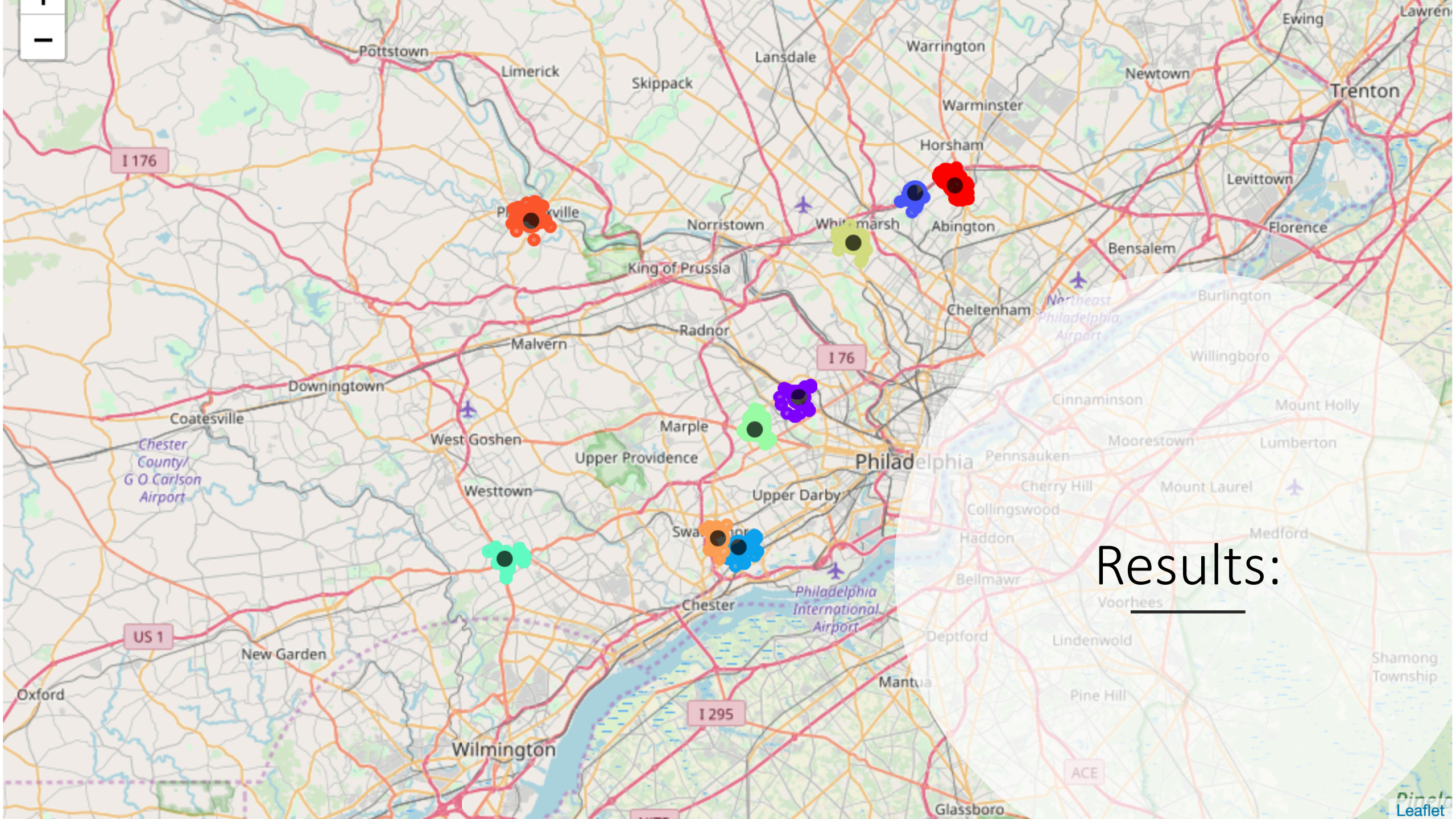
We ”score” each venue by taking the distance from the *edge*. This inverse distance means that the *lower* the number, the further away these venues are. And if the venue does not exist in the area (or if it's further away than 1 mile), then its inverse distance is 0. If there are several different venues of the same category, I simply average their inverse distances.

Methodology: Algorithm

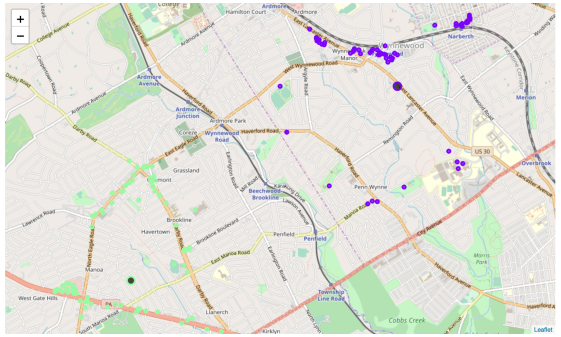
- To review, after setting up the data, each zip code is defined by a list of venue categories, with a number between 0 and 1600 representing the distance to the 'average location' of each venue category.

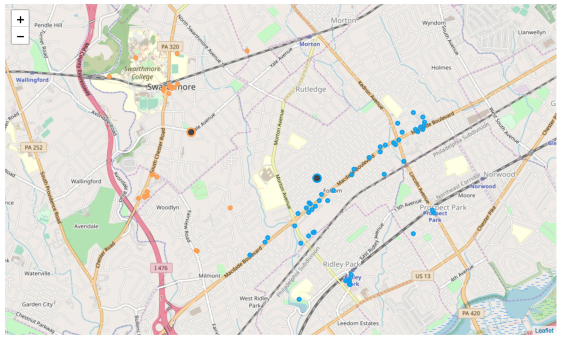
- We use an algorithm called K-Nearest-Neighbors (KNN) to identify the *k* zip codes that are most similar to our source zip code. Cosine similarity is used as the comparison metric because we don't care about absolute distances for each venue; rather, we want similar distance *relationships* with respect to the zip code center.

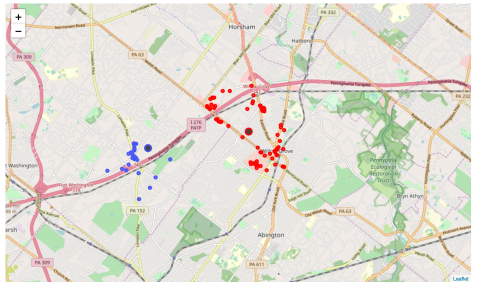



Pottstown
Limerick
Skippack
Lansdale
Warrington
Newtown
Trenton
Warminster
Horsham
Levittown
Florence
Bensalem
Abington
Norristown
King of Prussia
Cheltenham
Radnor
Malvern
Downingtown
Coatesville
Chester County/ G O Carlson Airport
West Goshen
Marple
Upper Providence
Westtown
Upper Darby
Philadelphia
Chester
Philadelphia International Airport
New Garden
Oxford
Wilmington
Glassboro
Burlington
Willingboro
Cinnaminson
Mount Holly
Moorestown
Lumberton
Pennsauken
Cherry Hill
Mount Laurel
Collingswood
Haddon
Medford
Bellmawr
Voorhees
Deptford
Lindenwold
Shamong Township
Pine Hill
I 176
I 295
US 1
ACE
Leaflet

Results:







Results & Discussion

- One of the major determining factors seems to be shopping centers - there are two within my current zip code detection radius, each of the similar zip codes also has a shopping center or shopping strip.
- When choosing the 10 most similar zip codes to my current one, the algorithm performs as expected. I am familiar with the local area, and all of the areas pinpointed on my map demonstrate a similar distribution of restaurants, shops, and services.

Conclusion & Next Steps

As I'm currently looking for a new place to live, this project was very timely! I learned an incredible amount about geographic mapping, hitting APIs to request data, and applying clustering techniques to understand similarities between groups.

Looking forward, given that the intent behind this project is to help find a new place to live, it would be very interesting to include cost-of-living information, as well as housing pricing and demographics information.

These additional data would make the comparison between zip codes more directly useful in terms of cost of living decisions.