Clustering of Neighborhoods in Washington D.C.

Plan

1. Introduction/Business Problem

1. **Problem:** Know which zip code in Washington D.C. share the same features.
2. **Goal:** Cluster zip codes that are similar in features.
3. **Target audiance:** This project will be useful for someone that wants to relocate in Washington D.C. The user will then have the option to chose which zip code has similar features as the one (s)he likes and is the closest to the point (s)he wants to relocate to.

2. Data

Our data comes from three different sources.
First Source: First, we obtained geographical coordinates of all zip codes in the US. We use this webpage: https://gist.github.com/erichurst/7882666
The table gives the GPS coordinates of each zip code

Second Source

The second source consist of all zip that are located in Washington D.C. We use this webpage: https://www.zillow.com/browse/homes/dc/district-of-columbia-county/
The table will give all zip codes in Washington D.C.

Third source

The third source of data is provided by the API of Foursquare. We use the API of Foursquare to fetch data related to venues around the coordinates of the zip code.

The Pre-processed and cleaned Data Set

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	20064	38.936354	-76.999167	&pizza	38.932582	-76.996696	Pizza Place
1	20064	38.936354	-76.999167	Chick-fil-A	38.935476	-76.998198	Food Service
2	20064	38.936354	-76.999167	Busboys and Poets	38.932117	-76.997640	American Restaurant
3	20064	38.936354	-76.999167	Starbucks Reserve	38.932484	-76.997172	Coffee Shop
4	20064	38.936354	-76.999167	BGR Burgers Grilled Right	38.932647	-76.996740	Burger Joint

3. Methodology

Which types of venues are the most common in Washington D.C. ?

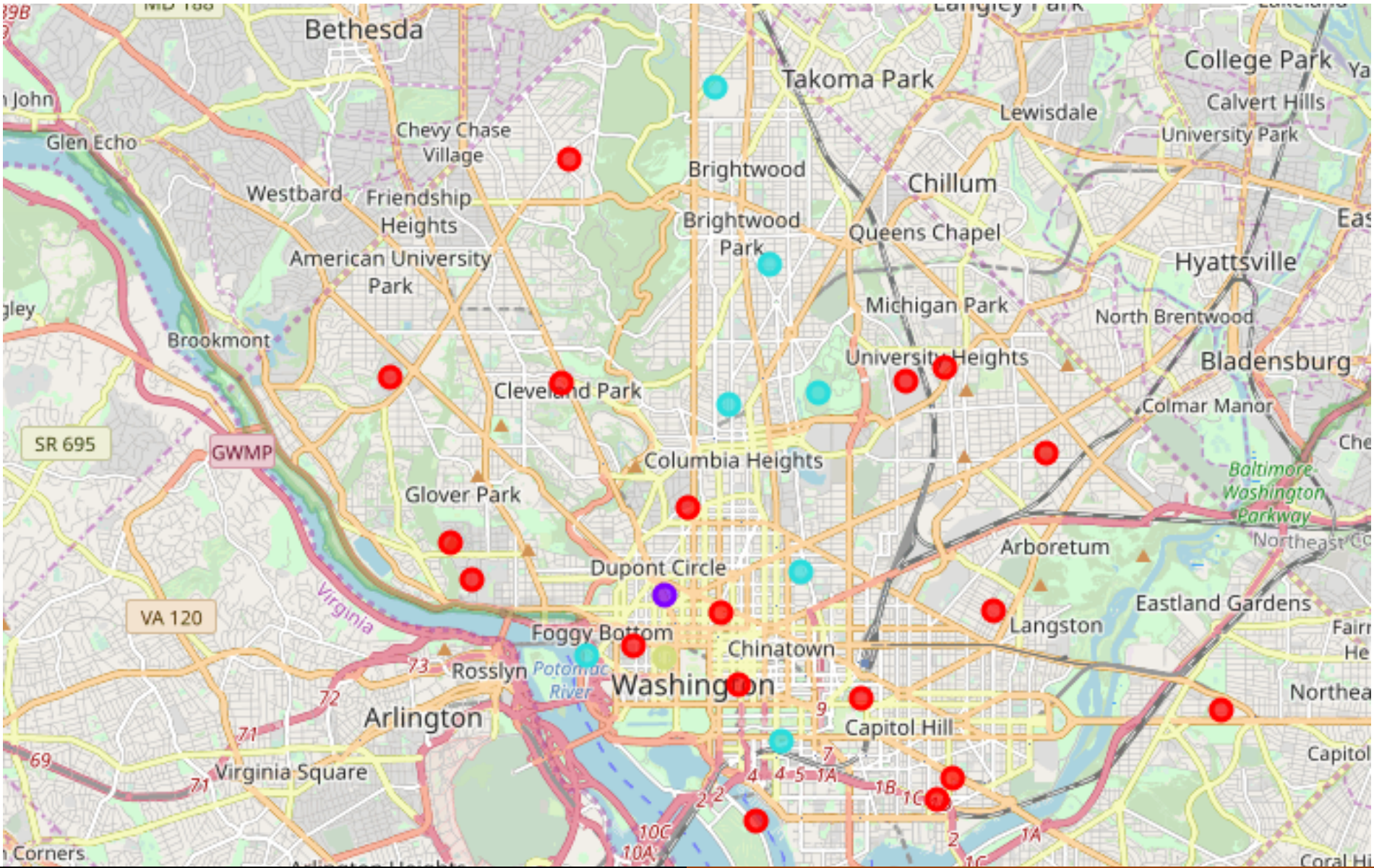
	Unnamed: 0	Frequency
0	Park	1.332066
1	Coffee Shop	1.071133
2	Sandwich Place	0.972823
3	American Restaurant	0.844335
4	Convenience Store	0.841558

We now figure out which venues are the most popular in each neighborhood.

	Unnamed: 0	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	0	20001	Thai Restaurant	Liquor Store	BBQ Joint	Grocery Store	Bookstore	Market	Building	Spanish Restaurant	Middle Eastern Restaurant	Gas Station
1	1	20002	American Restaurant	Bar	Gym	Moving Target	New American Restaurant	Park	Pharmacy	Diner	Convenience Store	Sandwich Place
2	2	20003	Pizza Place	Bar	Coffee Shop	Art Gallery	Gym / Fitness Center	Bakery	Sandwich Place	Spa	Mobile Phone Shop	Pet Store
3	3	20004	Hotel	Science Museum	History Museum	Exhibit	American Restaurant	Food Truck	Museum	Coffee Shop	Bakery	Sandwich Place
4	4	20005	Hotel	Hotel Bar	Coffee Shop	American Restaurant	Salon / Barbershop	Latin American Restaurant	Sandwich Place	Sushi Restaurant	New American Restaurant	Deli / Bodega

We decided to use a k-means clustering algorithm. We will be using the different venues as components of a 237 columns vector. After experimenting with different values of k, the number of clusters, we found that 4 clusters is optimal.

Machine-learning :Cluster map



4. Results

We were able throughout this project to cluster the neighborhoods in Washington D.C. in four clusters. Our final deliverable gives the option to the user to enter the zip code (s)he is moving from, the zip code (s)he is planning to move to. Then we recommend a zip code to the user based on which zip code is the nearest zip code to the target zip code that is also in of the same type (in the same cluster) as the original zip code.

Our clusters are as follows:

- Custer No 1: 20064, 20319, 20373, 20593, 20390, 20003, 20002, 20005, 20004, 20007, 20009, 20008, 20016, 20015, 20018, 20017, 20020, 20019, 20024, 20032, 20052, 20057, 20510
- Custer No 2: 20036
- Custer No 4: 20006
- Custer No 3: 20317, 20001, 20011, 20010, 20012, 20037, 20202

5. Discussion

It turns out that most neighborhood may be grouped in two groups. We have two neighborhoods that do have unique features. The first is the one around the World Bank. The second is the the neighborhood near Dupond Circle. The third cluster happens to be a band between Georgia Avenue and 16th Street. The fourth cluster, which is by far the largest, is all over the place.

6. Conclusion

Throughout this project, we obtained a segmentation of Washington D.C. in four clusters. This segmentation was based on most common venues in each neighborhood. Our projects gives the option to the user to enter the zip code (s)he is moving from, the zip code (s)he is planning to move to. Then we recommend a zip code to the user based on which zip code is the nearest zip code to the target zip code that is also in of the same type (in the same cluster) as the original zip code.

Thank you!