DC Capstone Project

By: Rory McLean

Objectives:

This project is the final step of the Capstone Project. In this project, location data for the Washington DC area will be used to (1) identify an ideal location for a new bike shop and (2) the ideal location for a new Asian cuisine restaurant. Neighborhoods will be defined by the latitudinal and longitudinal center of DC zip codes.

Analysis #1 For the bike shop analysis, the goal is to find an under-served area identified as:

- An area should have a large number of parks and bike trails.
- However, the area should be underrepresented by other bike shops and bike rental locations.

Analysis #2 For the restaurant analysis, the goal is to find an under-served area identified as:

- Areas with popular and successful restaurants as defined by the number of ratings (called "signals" by Foursquare).
- The idea areas is one that popular but does not have a highly rated restaurant in a given cuisine. For this analysis only Asian cuisine will be analyzed since ratings and signals require the use of a premium API which is limited by Foursquare. An average of all restaurants will be used for each zip code.

Data Sources:

The following data sources will be used:

- Areas will be identified by zip codes with the central point defined by the latitude and longitude. This information can be obtained from:
  http://www.city-data.com/city/Washington-District-of-Columbia.html
  and a file provided through a GIT site: https://gist.github.com/erichurst/7882666
- Foursquare will provide venues and reviews near the zip code latitude and longitude. For restaurants, the number of reviews ("signals") will denote the market potential and the rating will determine how competitive the venue will be. The data required is only available through the Foursquare premium API and due to lower daily limits with this API, analysis will be only be performed for Asian restaurants but this analysis could be used across all cuisines. For the bike shop, Foursquare will provide the number of bike shops in an area and the number of venues where a bike could be used, such as parks and trails.

Analytical Process:

Analysis #1 – Bike Shop Location

Methodology:

The analysis was limited to the DC zip codes as defined by this list:

> 20001, 20002, 20003, 20004, 20005, 20006, 20007, 20008, 20009, 20010, 20016, 20024, 20036, 20037, 20045, 20052, 20053, 20057, 20064, 20202, 20204, 20228, 20230, 20240, 20245, 20260, 20317, 20390, 20405, 20418, 20427, 20506, and 20510

The latitude and longitude for each zip code as submitted to Foursquare to identify the number of venues under several different categories:

> "Bike Shops" and "Bike Rentals" which were consolidated as "Total Bike Outlets"
>
> "Bike Trails", "National Parks", and "Parks" were consolidated as "Total Bike Uses"

Table 1 reveals that within ½ mile of the zip code centers, there are 18 bike shops and 320 bike rental locations. In terms of bike uses, there are 11 trails, 44 National Parks (which, for DC, may include just a building run by the National Park Service), and 551 parks.

Table 1: Venue Totals

Zip Code	Bike Shops	Bike Rentals	Bike Trails	National Parks	Parks	Total Bike Outlets	Total Bike Uses
20001	0	6	0	0	20	**6**	**20**
20002	0	3	0	0	6	**3**	**6**
20003	1	5	1	0	7	**6**	**8**
20004	1	19	1	4	19	**20**	**24**
20005	2	23	2	0	23	**25**	**25**
20006	0	19	0	1	32	**19**	**33**
20007	0	0	0	0	6	**0**	**6**
20008	0	2	0	0	13	**2**	**13**
20009	2	16	0	0	15	**18**	**15**
20010	0	7	0	0	14	**7**	**14**
20016	0	0	0	0	4	**0**	**4**
20024	1	5	1	0	14	**6**	**15**
20036	0	20	1	1	25	**20**	**27**
20037	1	8	0	3	17	**9**	**20**
20045	1	19	1	2	18	**20**	**21**
20052	1	20	0	0	28	**21**	**28**
20053	0	5	0	2	20	**5**	**22**
20057	2	4	0	1	5	**6**	**6**
20064	0	2	0	0	1	**2**	**1**
20202	0	8	0	2	20	**8**	**22**
20204	0	6	0	2	18	**6**	**20**
20228	0	5	0	3	12	**5**	**15**
20230	1	15	1	3	16	**16**	**20**

Zip Code	Bike Shops	Bike Rentals	Bike Trails	National Parks	Parks	Total Bike Outlets	Total Bike Uses
20240	0	11	0	3	25	**11**	**28**
20245	0	10	0	5	20	**10**	**25**
20260	1	10	1	1	17	**11**	**19**
20317	0	0	0	0	1	**0**	**1**
20390	1	4	1	0	7	**5**	**8**
20405	0	15	0	3	31	**15**	**34**
20418	0	7	0	6	18	**7**	**24**
20427	1	19	0	0	31	**20**	**31**
20506	1	18	1	1	30	**19**	**32**
20510	1	9	0	1	18	**10**	**19**
Total	**18**	**320**	**11**	**44**	**551**	**338**	**606**

An area deemed to be underserved is an area where "Total Bike Uses" far exceeds "Total Bike Outlets". The average ratio is 1.79 uses per bike outlet.

The leading six candidates for underservice are listed in Table 2:

Table 2: Leading Candidates

Zip Code	Bike Shops	Bike Rentals	Bike Trails	National Parks	Parks	Total Bike Outlets	Total Bike Uses	Ratio
20001	0	6	0	0	20	**6**	**20**	**3.33**
20008	0	2	0	0	13	**2**	**13**	**6.50**
20053	0	5	0	2	20	**5**	**22**	**4.40**
20204	0	6	0	2	18	**6**	**20**	**3.33**
20228	0	5	0	3	12	**5**	**15**	**3.00**
20418	0	7	0	6	18	**7**	**24**	**3.43**

From a graphical perspective (Figure 1) you can see the strong existing relationship between bike outlets and bike uses. Our interests lie in the points furthest below the line and ideally as far right as possible. The two largest circles in the graph are those two zip codes highlighted in Table 2.

An interactive graph is available at the following link: Biking Opportunities

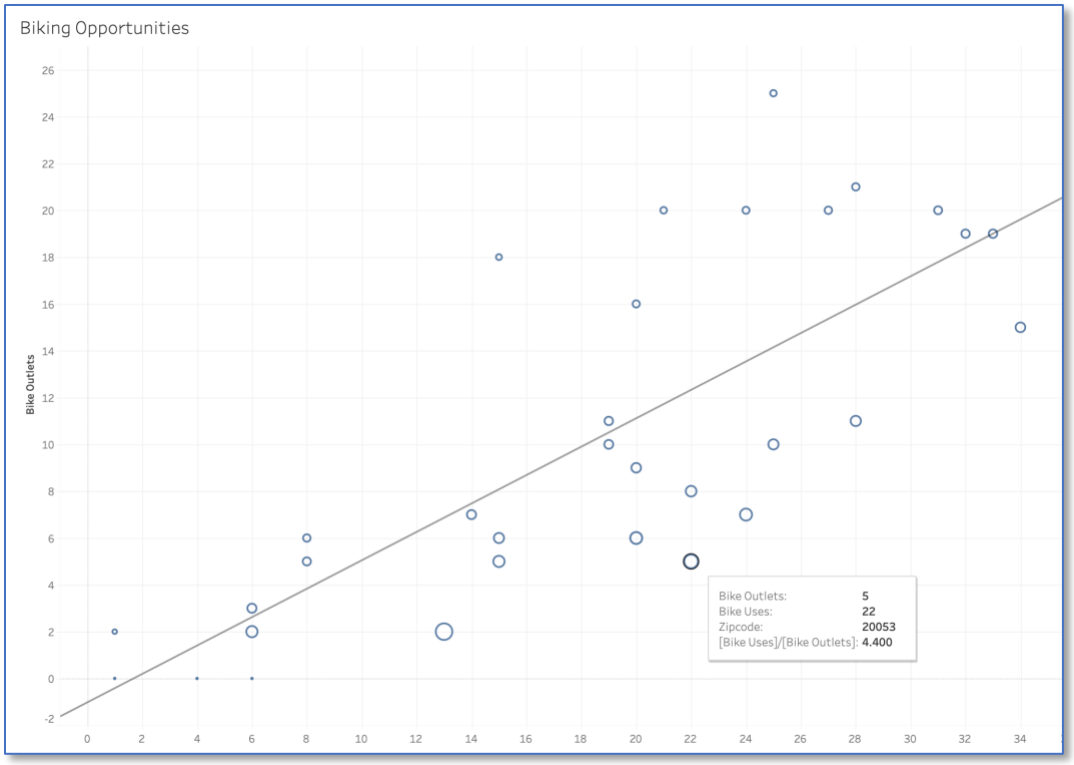

Figure 1: Relationship Between Outlets and Uses

A map indicates the relative strength of our two selected candidates.

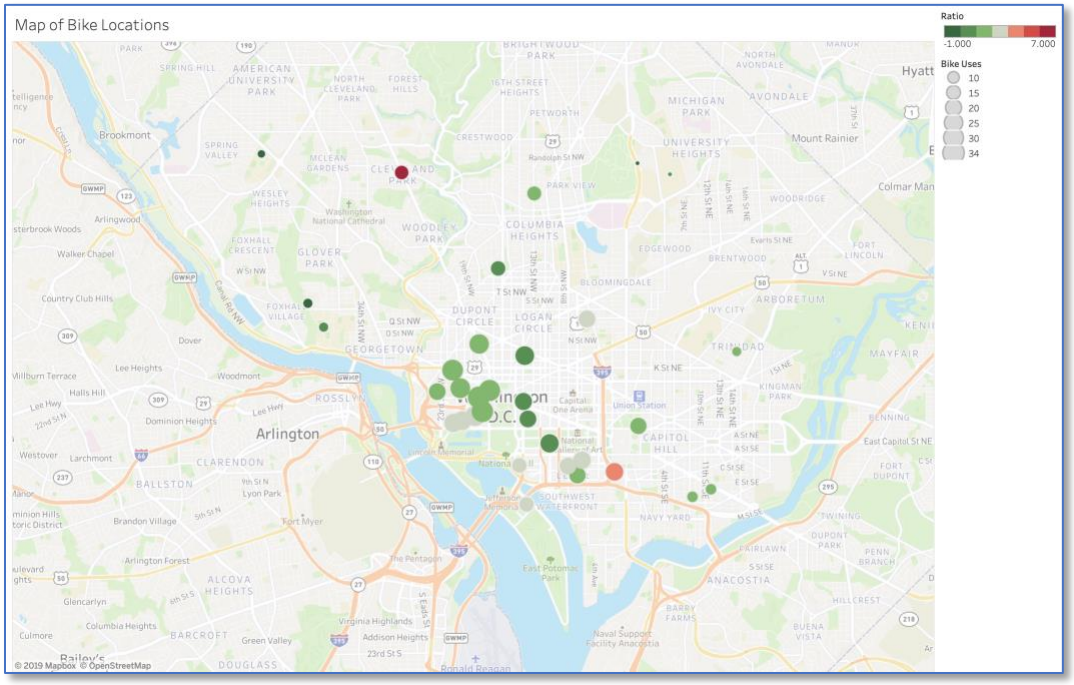

Figure 2: Map of Biking Opportunities

The red dot in Cleveland Park (20008) is in a more residential neighborhood. As the color indicates, the ratio makes this location one of the most underserved locations, but it lacks some of

the biking opportunities available in other zip codes. The orange dot (20053), located south of the Capital, is underserved to a lesser degree but offers more opportunities.

An interactive map is available at this link: Biking Opportunity Map

Recommendation:

As a marketing recommendation, both opportunities should be further explored. Keep in mind that 20053 is more of a tourist area than 20008 which could change the customer base and desired services.

Analysis #2 – Asian Restaurant Location

Methodology:

A similar analysis was done for Asian cuisine restaurants. However, with this analysis venue details were extracted and used to judge the relative strength of a location. An ideal site must have a high number of "signals", which is a Foursquare metric that includes ratings and tips for a venue. A high number of signals indicates the venue has significant traffic.

Ratings were used to measure if the area is underserved since low ratings but high signals indicated the demand is high but quality is low and there may be an opportunity for introducing an upscaled restaurant.

In the analysis, five candidates were selected based upon these criteria. Because of their low ranking, two candidates have been highlighted (20005 and 20427):

Table 3: Restaurant Candidates

Zip Code	Venues	Signals	Av Rating	Avg Signals
20004	11	552	7.791848	50.181818
20005	**16**	**977**	**7.554759**	**61.062500**
20006	16	627	7.825518	39.187500
20052	21	684	7.844444	32.571429
20427	24	1005	7.657811	41.875000

From a geographical perspective (Figure 3), the size of the dot indicates they are not the area with the most restaurants (which appears to be in the Dupont Circle area), but they are one of the larger areas and the color indicates that they underserved with and average score of only around 7.6.

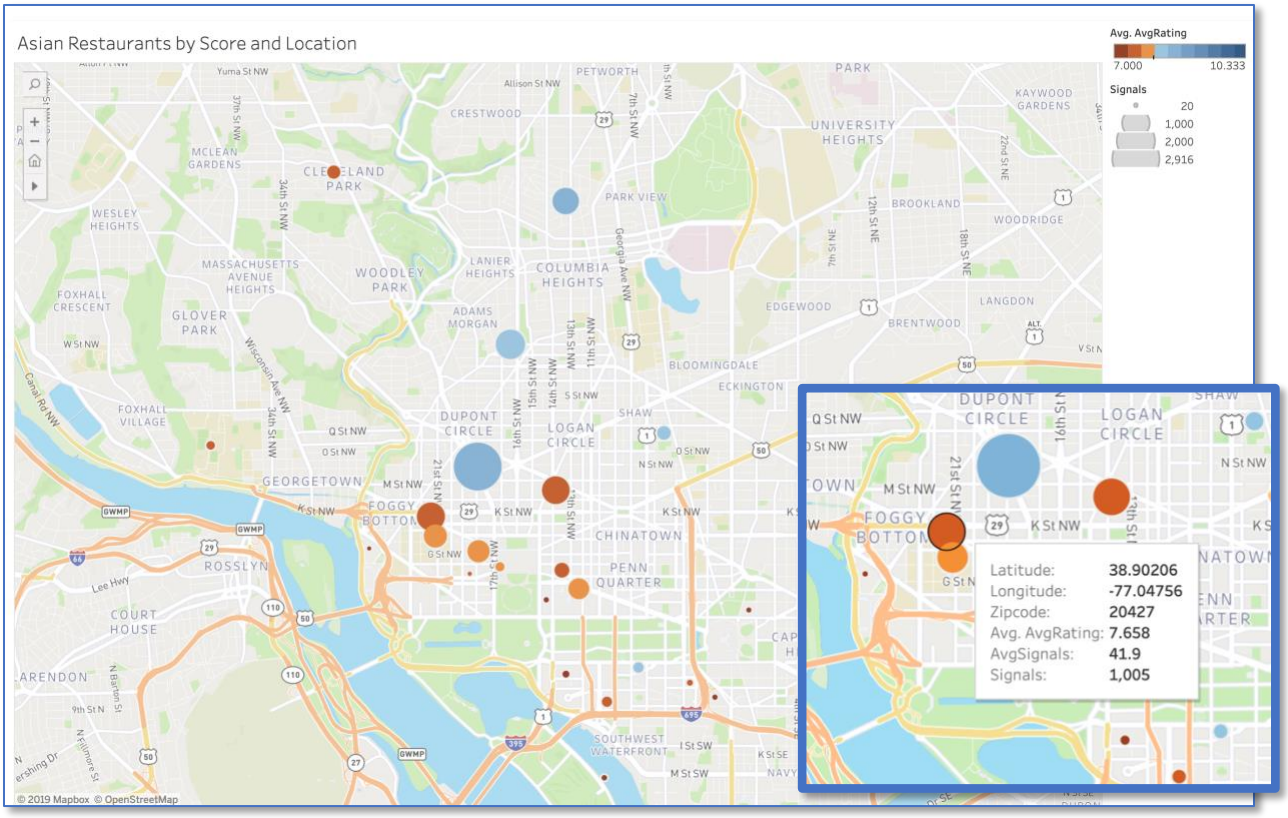

Figure 3: Restaurant Mapping

An interactive version of this map is available at this link: [Restaurant Map]

Recommendation:

Our recommendation is zip code 20005. This zip code has fewer restaurants than 20427 but nearly as many signals and a lower rating.

Area of Concerns and Improvements

DC is a particularly challenging due to its makeup. As the national capital, the center of the city is dedicated to the Federal Government. The large number of office buildings brings in workers and a large number of tourists. This changes demand for both bike shops and restaurants in ways that make it difficult to measure using the available data.

Zip code data may not be the best approach for this analysis since some area codes can be administrative districts comprising a few hundred square meters while other zip codes can be 100 square kilometers. A better approach would be to take all venues with specific latitudes and longitudes and performing a cluster analysis.

Lastly, Foursquare data lakes robustness for this data. Several restaurants I know are popular were not represented in the database. For the restaurant analysis it was necessary to use a premium API, which have limited quotas that I exceeded several times during this study.