**Neighborhood Recommendation Tool**

**Applied Data Science Capstone by IBM/Coursera**

**The Battle of the Neighborhoods**

## 1. Introduction

Problem Statement: A user is moving to a new city, and wants to better understand how each neighborhood in that city aligns with the user's particular taste/preference.

Introduction: Using foursquare API venue ratings, each neighborhood in a city is given a score for a set of neighborhood features. The final output will be a map which contains the neighborhood rankings for the user to explore.

For the sake of this exercise, the city will be Washington, DC, and the neighborhood features will be: food, drinks, shops, arts, and outdoors venues.

The target audience would be a prospective resident to Washington, DC.

## 2. Data

Based on the business problem, the following datasets are needed before performing the Content-Based recommendation algorithm to recommend a neighborhood:

1. **Washington, DC neighborhoods with latitude, longitude**

Approach: Use beautifulsoup to scrape neighborhood data from the DC.gov website (details are in the Jupyter notebook)

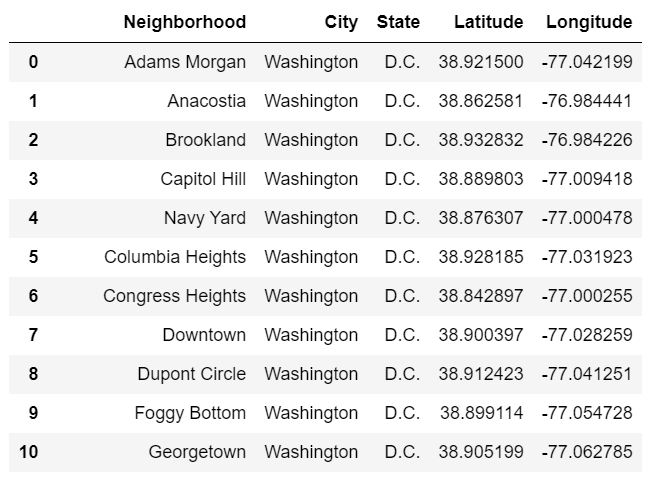


Figure 1. Snapshot of Washington DC Data.

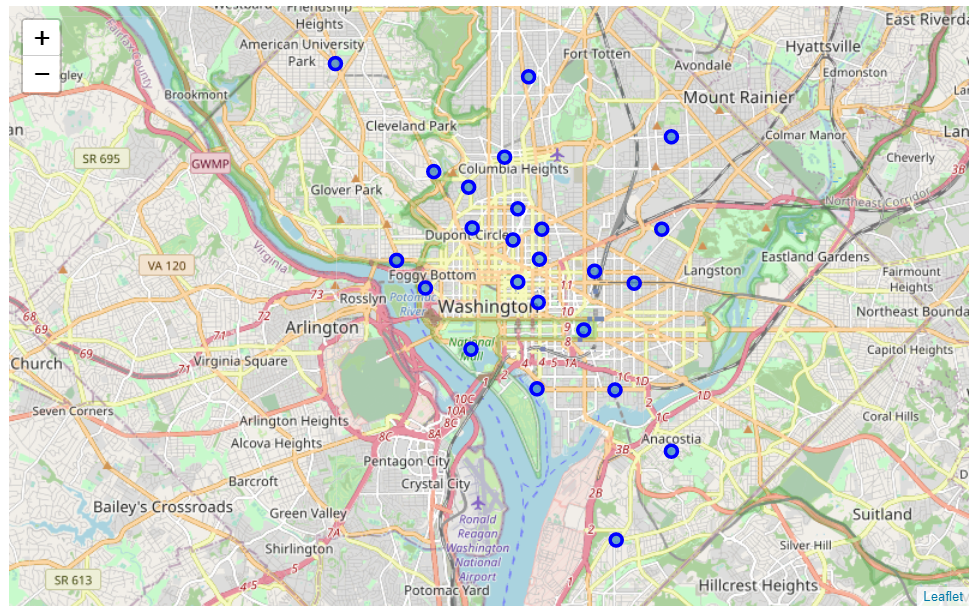


Figure 2. Map of DC neighborhoods.

1. **Foursquare venue type rating data associated with Washington, DC neighborhood**

Approach: define a http get function which queries foursquare’s API and returns a list of venues for each neighborhood. (details are in the Jupyter notebook)

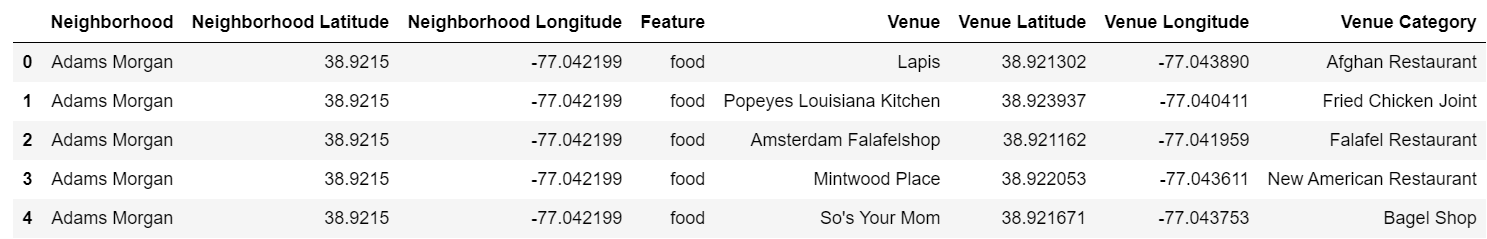


Figure 3. Snapshot of foursquare venue data.

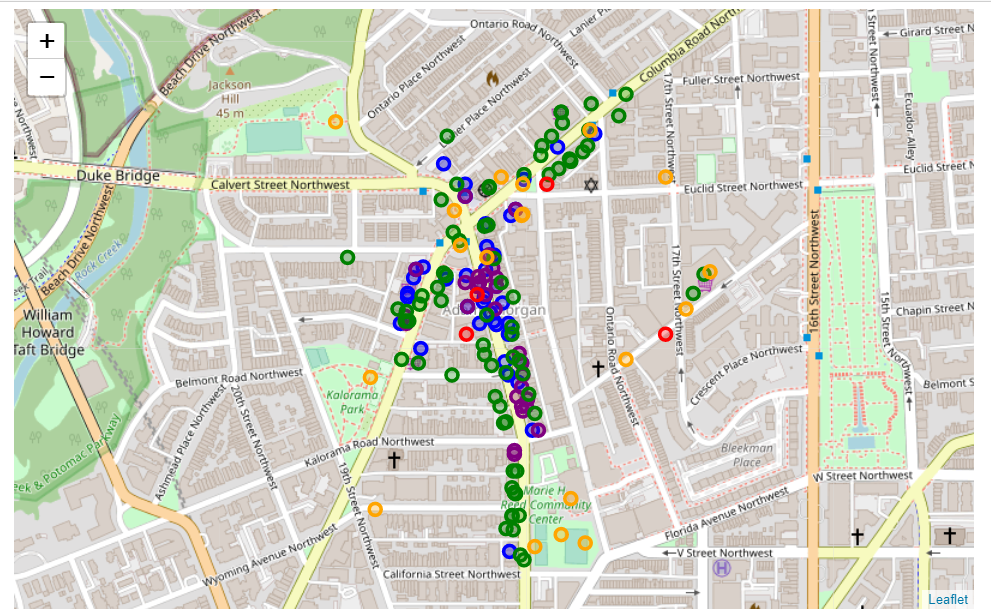


Figure 4. Example neighborhood’s venue data mapped.

1. User input rating of neighborhood features food, drinks, shops, arts, and outdoors venues

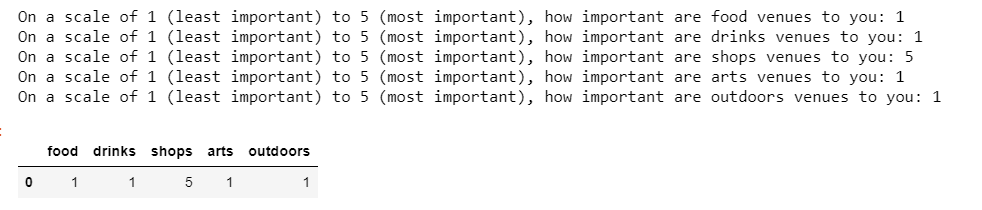


Figure 5. Snapshot of User Input process.

## 3. Methodology

In this project we will direct our efforts on identifying neighborhoods of Washington DC that have venue characteristics which align to the user's preference.

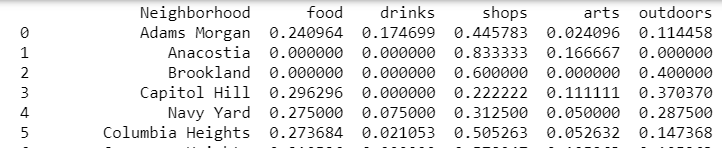
In first step we have collected the required data: neighborhood location and type (category) every venue within 400m of the neighborhood center.

The next step will be to normalize this neighborhood venue data to draw feature insights for each Washington DC neighborhood. For the sake of this exercise, the **number** of venues in each type will be used as the indicator of that type's prevalence in the area (versus something like average rating/proximity to the center).

In third and final step we will bring in the user's preferences to see which neighborhood is the best fit based on the preferences, and display that neighborhood with the venues for the user to explore.

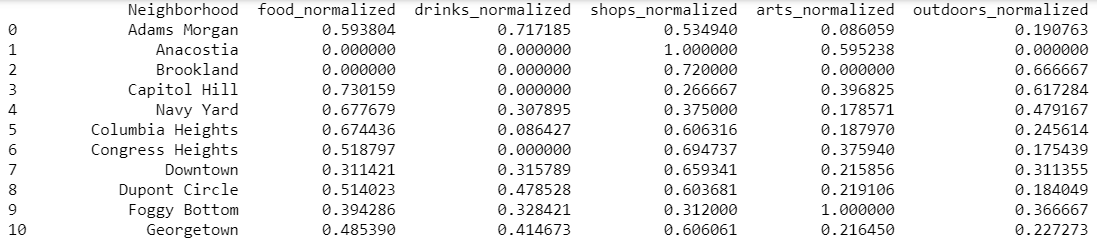
## 4. Analysis

First, I took the percentage of venue types in each neighborhood:



E.g. Of venues in Adams Morgan, 24% are food, 17% are drinks, etc.

Then, I compared these across neighborhoods to see how each neighborhood makeup compares across the city:



E.g. Adams Morgan is in the 59th percentile for food venues, 71st percentile for drinks, etc.

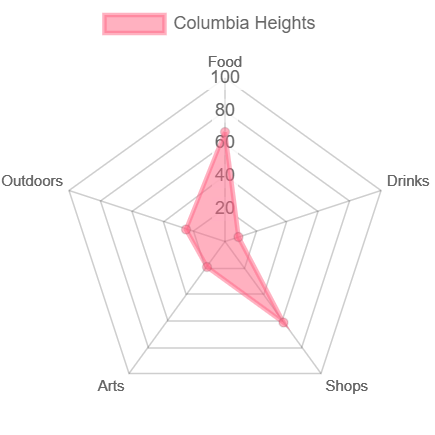
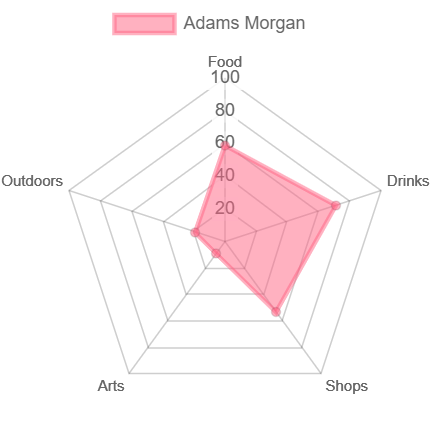


Figure 6. Adams Morgan versus Columbia Heights.

Then, a simple weighted average by the user’s ratings will yield the score for each neighborhood:

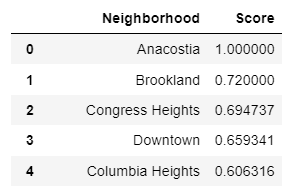
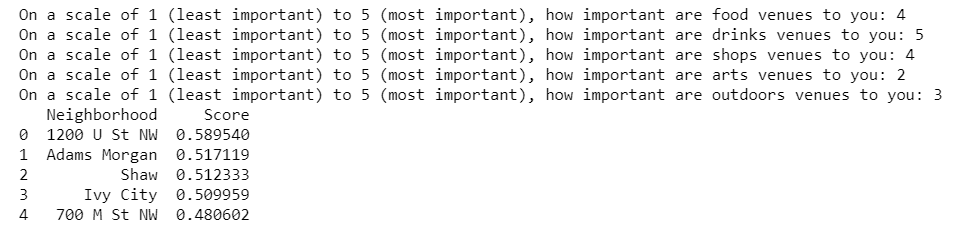


Figure 7. Snapshot of final recommendation dataframe.

## 5. Results and Discussion

A person who is moving into Washington, DC who prioritizes drinks venues the most, food and shopping second, outdoors third and arts last should ideally move into the U Street area, followed closely by the Adams Morgan and Shaw areas. This can be seen below:



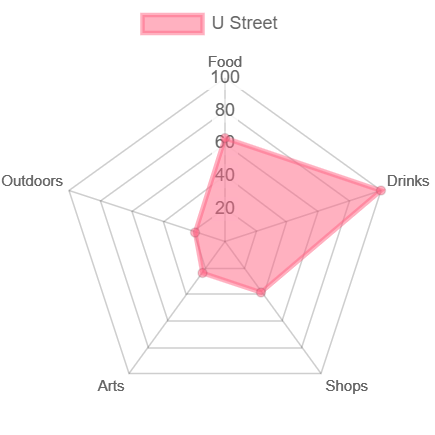


Figure 8. U Street Radar map.

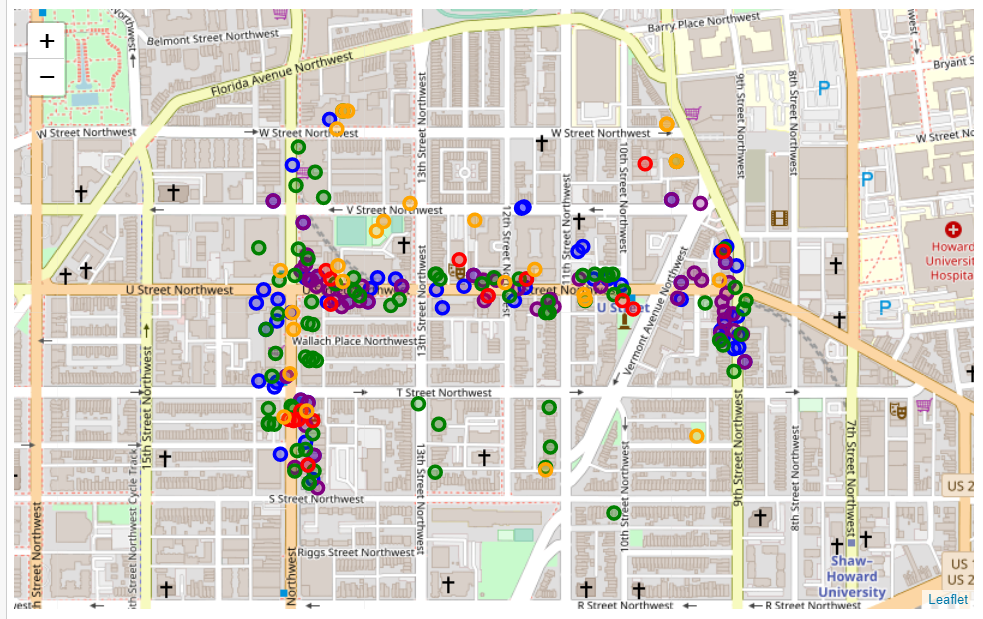


Figure 9. Map of the recommended area for the user to explore.

**Observations**

While performing the exploratory analysis on the DC neighborhoods, I had the following observations:

1. Anacostia, Brookland, and National Mall areas all had 10 or less venues when pulled from foursquare, which could skew the algorithm.



1. Minor issues with the foursquare data. Some foursquare categorizations of the venues could be off, as a restaurant can also be a place to get drinks, etc. Also, some venues come up multiple times due to multiple entries in foursquare.
2. The neighborhood vicinities were taken based off of radius from the center. In reality, neighborhoods are not perfect circles, but rather they are imperfect shapes. Ideally, venues should be considered which fall out of the radius but are realistically in the neighborhood.

## 6. Conclusion

Despite the difficulties presented in the Discussion section, the algorithm/model and functions serve the purpose proposed in the business problem. As a resident of Washington DC for the past 2 years, I feel the ratings which my algorithm came up with are fairly accurate with what I know of the neighborhoods in the area. It was particularly interesting to me that the algorithm recommended me to live in the U Street Area instead of my actual neighborhood, Adams Morgan.

I would advise any prospective resident to the area to play around with the user rankings and see the different neighborhoods it recommends; it would be a great first step in finding 2-3 neighborhoods to continue researching.