Recommending the best location for opening a restaurant in Washington DC

Pamal Meyana

December 24, 2020

Introduction

Background

Washington DC, being nation's capital attracts tourists from all over the world. There are several events that brings thousands of people every year to the city. My client "PastaPlace" plans to open a fast food restaurant in DC. They have a unique business model where all the menu items are ordered to go, and the customers can either use the dining space available in the restaurant or take the food to go. They are looking to open a new branch in the heart of Washington DC and would like to know the most popular location for restaurants in DC, where most people have checked-in and dined.

Objective

There are several popular restaurants in DC and there is no data readily available that provides the information on a popular location among tourists for restaurants. This project's goal is to collect the location information from USPS or similar Geodata provider and use the crowdsourced venue information from Foursquare to generate a popularity index and recommend the popular locations for opening a new restaurants.

Data Acquisition and Cleaning

Data Sources

The postal codes and corresponding latitude/longitude information is fetched from Opendatasoft, a French company that offers data sharing software. The API provides United States postal codes and lat/long in a JSON response. This data is saved into a "washington_dc_zip" local JSON file for processing.

```
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   datasetid             276 non-null    object
 1   recordid              276 non-null    object
 2   record_timestamp      276 non-null    object
 3   fields.city           276 non-null    object
 4   fields.zip            276 non-null    object
 5   fields.dst            276 non-null    int64
 6   fields.geopoint       276 non-null    object
 7   fields.longitude      276 non-null    float64
 8   fields.state          276 non-null    object
 9   fields.latitude       276 non-null    float64
 10  fields.timezone       276 non-null    int64
 11  geometry.type         276 non-null    object
 12  geometry.coordinates  276 non-null    object
```

The Places API from Foursquare is used to fetch the popular venue information within a 500 meter radius. The API returns a JSON response which includes, venue name, category, venue lat/long etc. Information from the two data sources are merged into a single Pandas data frame for statistical analysis and processing.

	zip	zip latitude	zip longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	20065	38.883412	-77.028198	ARTECHOUSE	38.884299	-77.029266	Art Gallery
1	20065	38.883412	-77.028198	Mandarin Oriental, Washington DC	38.883659	-77.030319	Hotel
2	20065	38.883412	-77.028198	International Spy Museum	38.883895	-77.025539	Museum
3	20065	38.883412	-77.028198	Maine Avenue Fish Market	38.881145	-77.028118	Fish Market
4	20065	38.883412	-77.028198	Falafel Inc	38.881443	-77.027500	Falafel Restaurant

Methodology

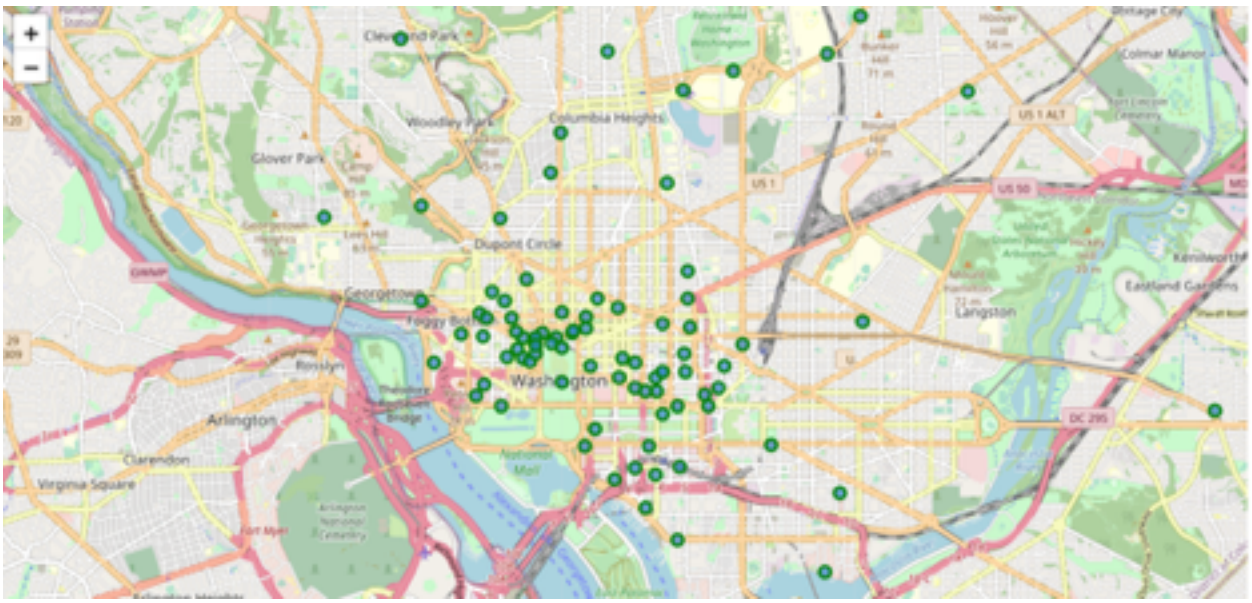
Data Analysis & Cleanup

The data returned by Opendatasoft contains several extra features that are not required for the modeling and analysis of this project. Hence all the irrelevant features are dropped from the dataframe. The data also contains multiple lat, long for the same zip code. Since the company PastaPlace would like to know the zip code information, all duplicate lat/long records are also deleted as part of data cleanup. The final data frame contains a total of 98 zip code records that are used for the modeling and recommendation.

	city	zip	long	lat
0	Washington	20210	-77.014647	38.893311
1	Washington	20202	-77.014647	38.893311
2	Washington	20537	-77.025133	38.894097
3	Washington	20068	-77.014647	38.893311
4	Washington	20076	-77.014647	38.893311

Inferential Testing

To get an initial idea of the spread of zip codes, the data frame is plotted on a map. The green dots on the map indicates the zip codes that are of interest in Washington DC.



The Foursquare venues categories are categorical information like “Asian Restaurant”, “Museum”, “Thai Restaurant” etc. This is converted to numerical indicator variables using Python get_dummies() function. We need the numerical values for performing computations on the data and identifying the popular areas. Each of the 98 zip codes have several venues in the area with the total number of unique venues as 278.

Item Based Recommendation

To find out the best places to open a restaurant, “item based recommendation” approach is used. The zip codes are grouped and the venue categories transposed to create a data frame.

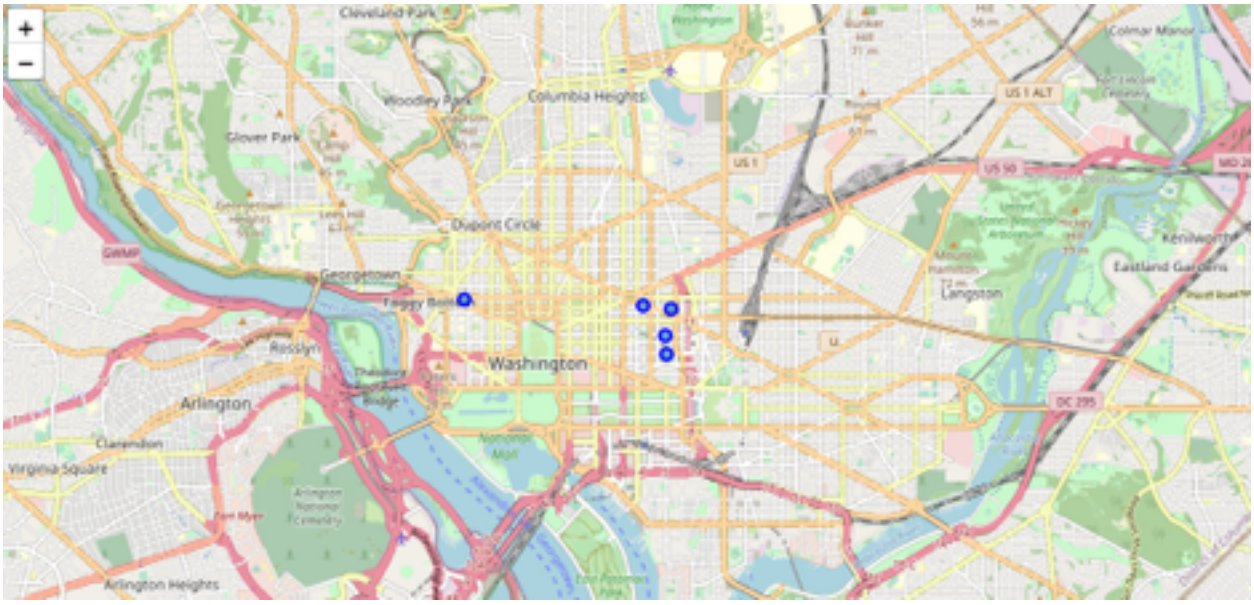
	zip	lat	long	ATM	Accessories Store	African Restaurant	American Restaurant	Arepa Restaurant	Art Gallery	Art Museum
0	20001	38.907711	-77.01732	0.0	0.0	0.0	0.00	0.00	0.00	0.00
1	20002	38.901811	-76.99097	0.0	0.0	0.0	0.02	0.00	0.02	0.00
2	20003	38.881762	-76.99447	0.0	0.0	0.0	0.04	0.00	0.04	0.00
3	20004	38.895268	-77.02760	0.0	0.0	0.0	0.04	0.00	0.00	0.04
4	20005	38.904461	-77.03088	0.0	0.0	0.0	0.04	0.02	0.00	0.00

Since this project's focus is on restaurants, all other features from the above data frame are dropped. There are a total of 52 different restaurant categories that will be used in the process. These zip codes are then assigned a score using the means() statistical function in Python.

	zip	lat	long	Score	African Restaurant	American Restaurant	Arepa Restaurant	Asian Restaurant	Belgian Restaurant	Brazilian Restaurant	Burmese Restaurant	Cajun / Creole Restaurant
0	20001	38.907711	-77.01732	0.142857	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	20002	38.901811	-76.99097	0.240000	0.0	0.02	0.00	0.04	0.00	0.00	0.02	0.02
2	20003	38.881762	-76.99447	0.340000	0.0	0.04	0.00	0.02	0.02	0.00	0.00	0.00
3	20004	38.895268	-77.02760	0.160000	0.0	0.04	0.00	0.00	0.00	0.02	0.00	0.00
4	20005	38.904461	-77.03088	0.280000	0.0	0.04	0.02	0.00	0.02	0.00	0.00	0.00

Results

The data frame with scores are then sorted in descending order to find the most popular venues. The top 5 zip codes are then plotted on a map using folium to show the recommended places to open a restaurant.



Following are the zip codes recommended to **PastaPlace** to open a restaurant in Washington DC

20548,
20586,
20055,
20442,
20536

Conclusion

This project report provides a list of recommended zip codes to start a new restaurant in Washington DC, based on the venue information available from Foursquare. This report can be further improved by adding a weighted score to the venues based on factors like “rental expense”, “most visited place”, “most liked” etc. in future. This will allow PastaPlace to get a more personalized recommendation based on their preferences.