Choosing the Best Location for a Café or Restaurant in Washington DC

Nathan Edgerton

April 30, 2020

1. Introduction

1.1 Background

Choosing the location to open up a new café or restaurant is an important decision for an entrepreneur. With typically low profit margins, it is important to choose a location in a neighborhood that is already saturated with similar businesses. Since most people tend to frequent cafes and restaurants nearby their homes, the most important competitors for a foodservice business are those that are nearest.

1.2 Problem

An entrepreneur interested in establishing a new foodservice business will first need to choose a location to rent. It would be extremely time-consuming to search for available locations to rent around the city and then to travel to each neighborhood to research the number of competing businesses nearby. Moreover, due to these high search costs, entrepreneurs may sign a contract to rent a location before thoroughly investigating a range of available options. Rental agreements typically require a long-term commitment, and thus a very significant fixed expense for an entrepreneur.

The entrepreneur faces the risk of choosing a location in an area with too many competitors, and they also face the risk of choosing a location without an area without sufficient population or disposable income to support the business.

1.3 Interest

Since the choice of location is so risky for an entrepreneur, they would certainly be interested in gaining some insight to narrow down their search for the best location. If they can tell which neighborhoods appear to be the least saturated with cafes and restaurants, they can save time in their search and feel more confident in their decision.

2. Data Acquisition and Cleaning

2.1 Data Sources

I obtained data from Open Data DC on average income and population for each census tract in Washington DC from the 2010 census data.[i] This data had been combined by a helpful Github user, benbalter, with the coordinates marking the boundaries of each census tract in a GeoJSON file.[ii] There are a total of 179 census tracts in this dataset.

I then obtained data on the restaurants and cafes in each neighborhood through the Foursquare API.

2.2 Data Preparation

Having obtained the coordinates of the border of each census tract, I next had to estimate the central point (centroid) of each census tract. I used an algorithm to estimate the centroid by averaging the minimum and maximum of the latitudes of the border points and the minimum and maximum of the longitudes of the border points to obtain the centroid. This worked well generally, though it didn't work so well for some irregularly-shaped tracts.

After getting the centroid of each census tract, I then used the Foursquare API to query and return a list of all venues listed within a 500-meter radius of each centroid. The resulting dataset had 4,002 venues and 321 unique categories of venues, including venues such as art galleries, museums, cafes, and restaurants. For each census tract, I summed all venues categorized as 'café' and coffee shop' into a category called 'cafes.' I then summed all venues whose category included the word 'restaurant' into a category called 'restaurants.'

After merging the data set containing median income and population data with the data set containing data on the number of cafes and restaurants within 500 meters of the centroid of each tract, I obtained the following data frame (showing only the first five rows):

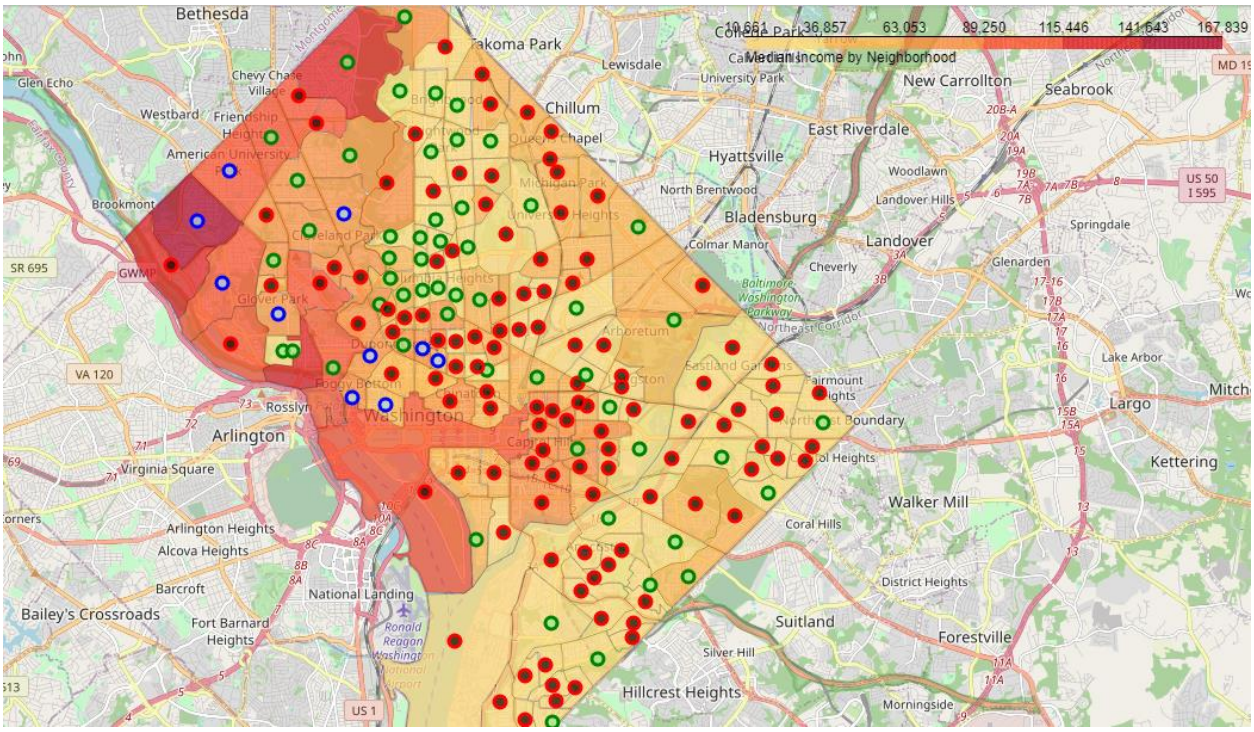
	tract_id	total_pop	median_income	long	lat	Total_cafes	Total_restaurants
0	001001	7436	114136.5	-77.089557	38.949481	0	0
1	001002	3442	74658.0	-77.079024	38.939686	2	3
2	004001	3745	72807.0	-77.046452	38.919678	7	18
3	004002	2797	60460.5	-77.043998	38.918528	6	27
4	004100	2708	87019.0	-77.052629	38.915475	0	1

3. Exploratory Data Analysis

3.1 Visualization of Median Income and Population

I first created a chloropleth map to investigate the distribution of median income around the city. From the visualization, where dark red indicates high median income and lighter colors indicate low median income, we can see that the highest incomes are concentrated in the northwest, west and center of the city.

I also plotted circle markers on the map to show the total population of each census tract. A red marker shows that the tract has a population under 3500, a green marker shows a population between 3500 and 6000, and a blue marker shows a population over 6000. All else equal, entrepreneurs would tend to prefer to locate in areas with a high population and high median income, so tracts in the center and northwest of the city would appear to be most attractive at this stage.



3.2 Data Selection

I next plotted the number of cafes and restaurants in each census tract against the tract's median income and population.

The first point to consider is that the graphs show there are many census tracts with both zero cafes and zero restaurants, or only one of each. Considering the method for choosing the centroid of each census tract, there was a risk that the algorithm could place the centroid in the middle of a park or in some other area where, if a 500 meter radius were drawn around that point, there would be few cafes or restaurants observed. Because of this, I decided to drop census tracts where the total number of cafes and restaurants was less than or equal to 2, taking this low number to be evidence of either a misplace centroid, or a census tract that was purely residential due to zoning restrictions. After dropping these census tracts, the total number of tracts decreased from 173 to 92.

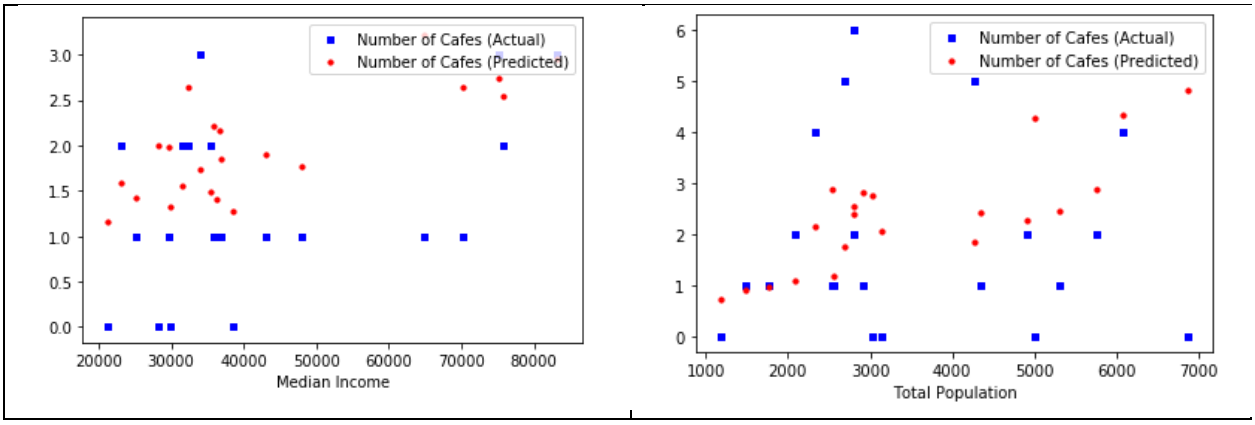
Next, considering the graph showing median income, I observed that there were a few outlier communities at the higher end of the distribution that had very few or no cafes or restaurants. I took this anomalous result as evidence that the community is primarily residential. Concerned that these few observations could bias the regression results, I dropped communities with a median income above $100,000. After doing this, the total number of census tracts decreased from 92 to 90.

4. Predictive Modeling

4.1 Predicting the Number of Cafes in Each Census Tract

The goal of the study was to predict the number of cafes and restaurants in each neighborhood based on the median income and population of that neighborhood. Using these predicted values, we can then see for which neighborhoods the actual number of cafes or restaurants is lower than the predicted number by the largest amount. These may be the most promising neighborhoods in which to investigate setting up a new foodservice business. As these are all continuous variables, I decided to use multiple linear regression as the predictive model.

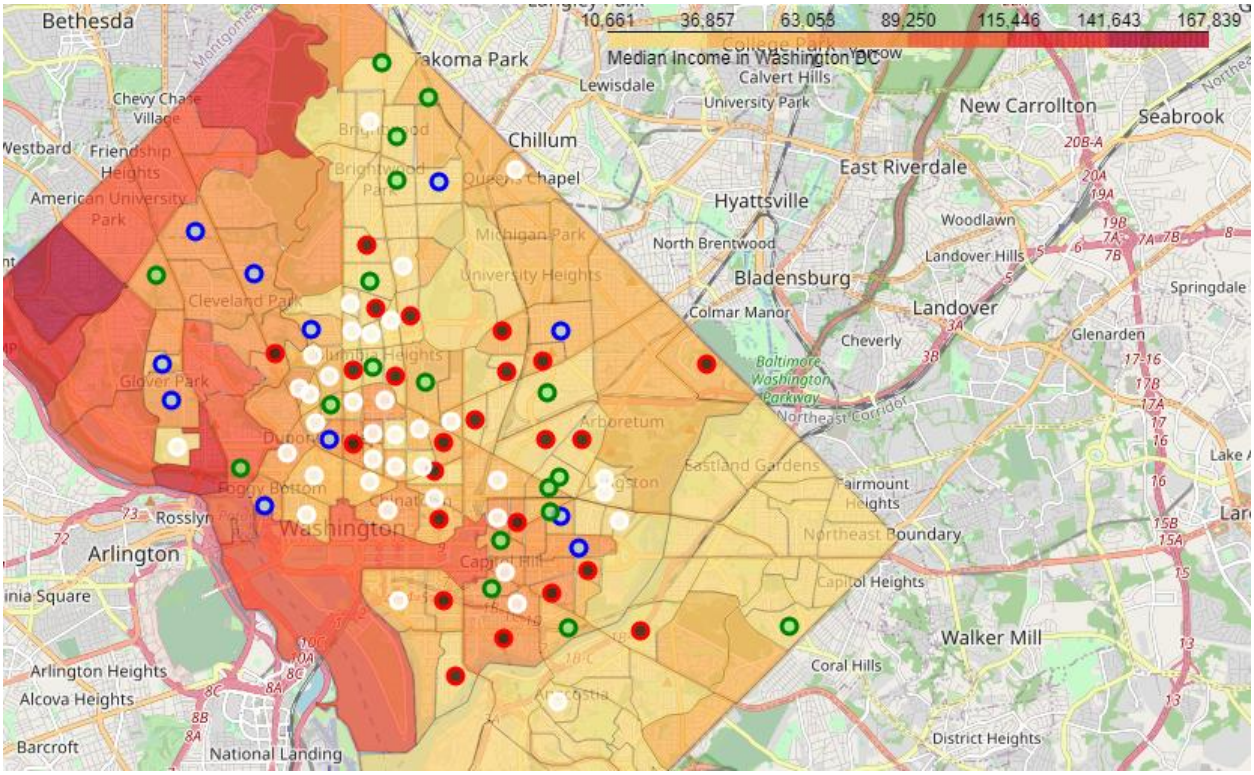
I first ran a regression on a training set of data (using 69 samples) with the median income and the population as independent variables and the number of cafes as the dependent variable. I then used the regression model to estimate the number of cafes in the test set (21 samples). I then plotted the predicted number of cafes for the test set against the actual number of cafes, using both median income and total population on the x-axis.



As we can see, the variance of the predicted values is much lower than the variance of the actual values. This indicates that there may be other important influences on the number of cafes in each census tract. This might include data such as proximity to public transportation, presence of shopping centers or malls which could lead to a clustering of foodservice businesses, or proximity to office buildings. The R2 score of this regression was -3.14, which indicates a poor fit of the model.

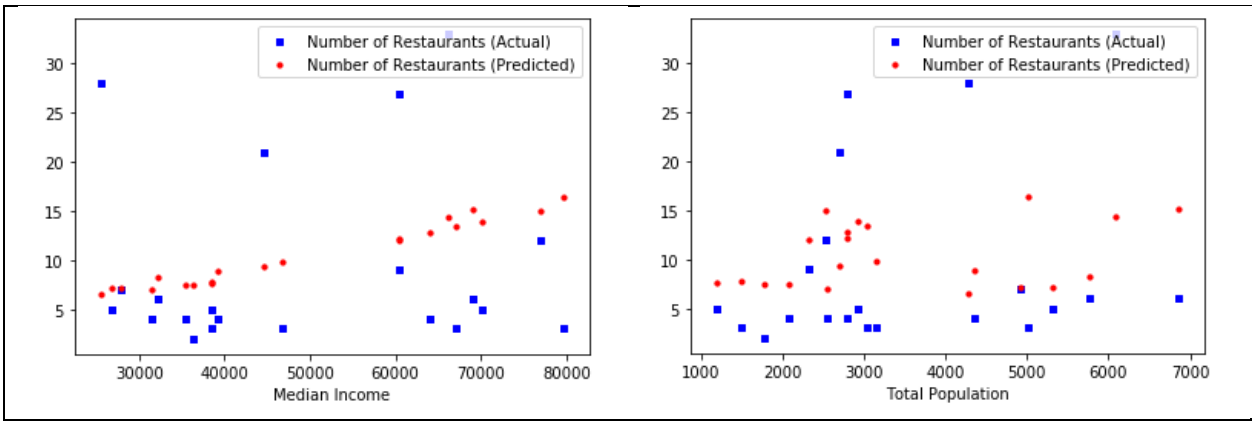
Despite the low accuracy of the predictive model, I then predicted the number of cafes for the entire dataset and then calculated the difference between the number of predicted cafes and actual cafes. I next sorted the dataset to focus on the neighborhoods which had the largest difference to show which neighborhoods appear to be the best prospects for opening a new cafe. Because of the low accuracy of the regression model, though, these recommendations would have to be taken with a big grain of salt.

On the map below, blue markers indicate that a neighborhood has at least 2 fewer cafes than predicted by the model. A green marker indicates between 1 and 2 fewer cafes, and a red marker indicates between 0 and 1 fewer cafes. A white marker indicates that the neighborhood has more cafes than predicted by the model.



4.2 Predicting the Number of Restaurants in Each Census Tract

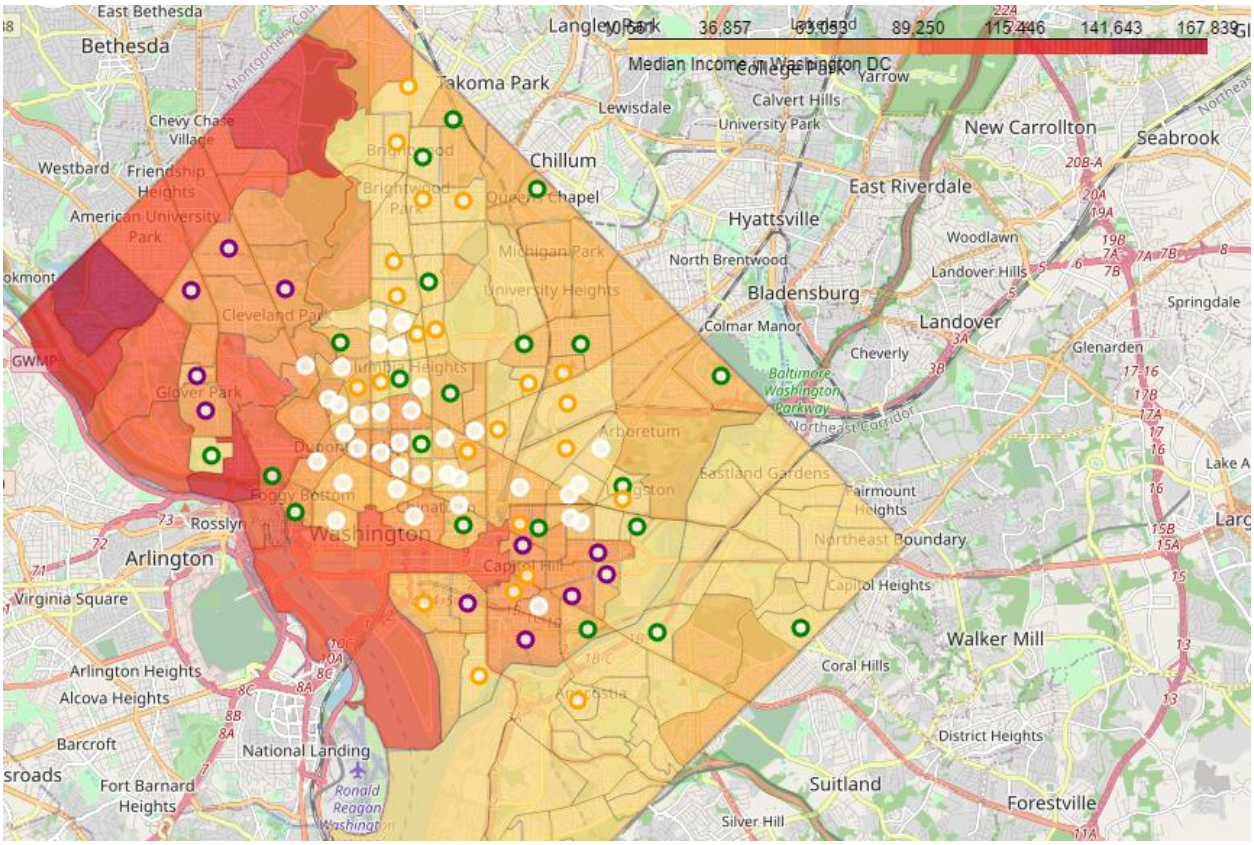
I next ran the same regression as above, except with the number of restaurants as the dependent variable. This regression encountered the same problem as the regression above, with a negative R2 value of -7.73.



Despite the low accuracy of the predictive model, I then predicted the number of restaurants for the entire dataset and then calculated the difference between the number of predicted restaurants and

actual restaurants. I next sorted the dataset to focus on the neighborhoods which had the largest difference to show which neighborhoods appear to be the best prospects for opening a new restaurant. As above, the recommendations must be interpreted cautiously.

On the map below, purple markers indicate that a neighborhood has at least 7 fewer restaurants than predicted by the model. A green marker indicates between 3 and 7 fewer restaurants, and an orange marker indicates between 0 and 3 fewer restaurants. A white marker indicates that the neighborhood has more restaurants than predicted by the model.



5. Conclusion

In this study, I attempted to use median income and population to predict the number of cafes and restaurants that can be supported within each census tract in Washington DC. The purpose was to identify which census tracts had a lower than predicted number of cafes or restaurants, as these could be ideal sites for entrepreneurs to consider if they would like to open a new foodservice business.

Due to the low predictive accuracy of the regression models, the results must be interpreted cautiously. However, I hope that the data visualizations generated could be useful to entrepreneurs who are interested in where to begin their search for a location to open a new business.

6 Further Directions

This study could be improved in a number of ways, including by adding more data on possible relevant factors, such as proximity to office buildings, tourist attractions, or other venues that attract many

customers, location of shopping centers and public transportation, and other factors that could impact the number of cafes or restaurants that a neighborhood could support.

Since census tracts vary significantly in size, it could be useful to expand the radius around the centroid from which venues are detected using the Foursquare API call. For example, in a smaller census tract perhaps a radius of 300 meters would be used, whereas in a larger census tract perhaps 600 meters would be used, in order to gather data on businesses from a higher percentage of the census tract area.

Finally, the analysis could be updated with 2020 Census Data once that becomes available.

---

[i] https://opendata.dc.gov/datasets/6969dd63c5cb4d6aa32f15effb8311f3_8

[ii] https://raw.githubusercontent.com/benbalter/dc-maps/master/maps/census-tracts-2010.geojson