Find a place to open a fast-food restaurant

Kwai Leung

09/28/2020

Contents

[Background 2](#_Toc52223851)

[Problem 2](#_Toc52223852)

[Data Acquisition and Cleaning 2](#_Toc52223853)

[Data Sources 2](#_Toc52223854)

[Data Cleaning 2](#_Toc52223855)

[Methodology/Exploratory Data Analysis 3](#_Toc52223856)

[First Analysist - 5 Clusters 4](#_Toc52223857)

[Population – Cluster may contain same city but different zip code 4](#_Toc52223858)

[Number of restaurants and portion of each cluster 5](#_Toc52223859)

[Second Analysist - 7 Clusters 5](#_Toc52223860)

[Population – Cluster may contain same city but different zip code 5](#_Toc52223861)

[Number of restaurants and portion of each cluster 7](#_Toc52223862)

[Conclusions 7](#_Toc52223863)

[Future directions 7](#_Toc52223864)

# Background

Maryland, US is one of the US States in Mid-Atlantic region and bordering Pennsylvania, Virginia, and Washington District of Columbia. Based on the United States Census Bureau, population of Maryland has almost 6 million in 2015 and keep growing in the past 20 years. According to the National Restaurant Association, restaurants are a driving force in Maryland’s economy, about 11,357 established eating and drinking places in 2018. Estimated $13.3 billion sales in Maryland’s restaurant in the same year. According to the Washington Examiner newspaper’s article on June 06, 2015, 2.7767 fast-food restaurants per 10,000 people and it was ranked in 37 out of 50 US States. Another article on May 27, 2017 on the website, BrandonGaille Small Business & Marketing Advice, said Maryland was one of the top 10 States for fast-food consumption and it was ranked in number 4. Therefore, you can see a great potential for fast-food business in Maryland.

# Problem

A customer wants to open a fast-food restaurant in Maryland, US. Therefore, they asked us to help them and find an ideal location for the restaurant. To find the candidate locations, we will consider the following criteria and use the data science technique to resolve this problem. The candidate locations should meet all criteria below for analysis.

First, we want to find the city with high population which is at least 45,000 in this case.

Second, the city has low proportion of fast-food restaurant per people to minimize the competition.

# Data Acquisition and Cleaning

## Data Sources

To get the latitude, longitude, and estimated population data, we download the data in csv format [here](https://www.unitedstateszipcodes.org/zip-code-database/) from United States Zip Code organization. The data covers the entire United States. For the location and its surrounding restaurant data, we use Foursquare API to retrieve related data.

## Data Cleaning

The data from United State Zip Code organization covers entire United State, but all we want is Maryland data only. Therefore, we needed to filter the State to Maryland. Also, we removed the columns that we do not need and filtered the population greater or equal to 45,000 and saved it as a new data frame data. After that, we got 18 Maryland records.

At the beginning, we kept most of the columns from the United State Zip Code organization because we were not sure those columns would be useful later or not. On the other hand, we pulled the data from Foursquare and we separated the venue column plus neighborhood as a separate dataset, MD\_onehot. And grouped MD\_onehot by neighborhood and calculated the mean for each fast-food restaurant.

Finally, we merged both data from United State Zip Code organization and the grouped data from Foursquare, MD\_onehot, into a new data frame and removed the unnecessary columns for further analysis.

# Methodology/Exploratory Data Analysis

Now, we use one of the data science techniques, k-means clustering, to analysis the data and find the candidate locations. First, we set the number of clusters as 5 and perform analysis. Then we will set the number of clusters as 7 and perform second analysis. K-means clustering divides the data into non-overlapping clusters without any cluster internal structure. The fast-food restaurants within a cluster are similar in this case. Here is the map after applying 5 clusters and you can see 5 different colors there.

Map

Description automatically generated

Next, we are going to see the number of population and the number of fast-food restaurants in each cluster, then calculate the radio between them. Let us look at the result of each cluster below.

## First Analysist - 5 Clusters

### Population – Cluster may contain same city but different zip code

Cluster 1

|  |  |
| --- | --- |
| **City** | **Estimated Population 2015** |
| Parkville | 62620 |
| Pasadena | 57620 |
| Owings Mills | 52350 |
| Glen Burnie | 51090 |
| Dundalk | 50150 |
| Baltimore | 47150 |
| Baltimore | 45550 |
| Ellicott City | 45480 |
| **Total** | **412010** |

Cluster 2

|  |  |
| --- | --- |
| **City** | **Estimated Population 2015** |
| Gaithersburg | 62930 |
| Germantown | 59300 |
| **Total** | **122230** |

Cluster 3

|  |  |
| --- | --- |
| **City** | **Estimated Population 2015** |
| Silver Spring | 68290 |
| Silver Spring | 55730 |
| Silver Spring | 52380 |
| Hyattsville | 49520 |
| Potomac | 48030 |
| **Total** | **273950** |

Cluster 4

|  |  |
| --- | --- |
| **City** | **Estimated Population 2015** |
| Hagerstown | 52380 |
| **Total** | **52380** |

Cluster 5

|  |  |
| --- | --- |
| **City** | **Estimated Population 2015** |
| Fort Washington | 50410 |
| Upper Marlboro | 45540 |
| **Total** | **95950** |

### Number of restaurants and portion of each cluster

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster** | **Total Population** | **Number of fast-food restaurants** | **Ratio of fast-food restaurant and population** |
| 1 | 412010 | 473 | 871 |
| 2 | 122230 | 100 | 1222 |
| 3 | 273950 | 271 | 1010 |
| 4 | 52380 | 39 | 1343 |
| 5 | 95950 | 103 | 931 |

As you can see, cluster 4 has the highest ratio of fast-food restaurant and population. Now, let us examine the result when the number of clusters is 7. Here is the map of the distribution.

Map

Description automatically generated

## Second Analysist - 7 Clusters

### Population – Cluster may contain same city but different zip code

Cluster 1

|  |  |
| --- | --- |
| **City** | **Estimated Population 2015** |
| Parkville | 62620 |
| Dundalk | 50150 |
| Baltimore | 47150 |
| Baltimore | 45550 |
| **Total** | **205470** |

Cluster 2

|  |  |
| --- | --- |
| **City** | **Estimated Population 2015** |
| Hagerstown | 52380 |
| **Total** | **52380** |

Cluster 3

|  |  |
| --- | --- |
| **City** | **Estimated Population 2015** |
| Silver Spring | 68290 |
| Silver Spring | 55730 |
| Silver Spring | 52380 |
| Hyattsville | 49520 |
| Potomac | 48030 |
| **Total** | **273950** |

Cluster 4

|  |  |
| --- | --- |
| **City** | **Estimated Population 2015** |
| Fort Washington | 50410 |
| Upper Marlboro | 45540 |
| **Total** | **95950** |

Cluster 5

|  |  |
| --- | --- |
| **City** | **Estimated Population 2015** |
| Gaithersburg | 62930 |
| Germantown | 59300 |
| **Total** | **122230** |

Cluster 6

|  |  |
| --- | --- |
| **City** | **Estimated Population 2015** |
| Pasadena | 57620 |
| Glen Burnie | 51090 |
| **Total** | **108710** |

Cluster 7

|  |  |
| --- | --- |
| **City** | **Estimated Population 2015** |
| Owings Mills | 52350 |
| Ellicott City | 45480 |
| **Total** | **97830** |

### Number of restaurants and portion of each cluster

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster** | **Total Population** | **Number of fast-food restaurants** | **Ratio of fast-food restaurant and population** |
| 1 | 205470 | 245 | 838 |
| 2 | 52380 | 39 | 1343 |
| 3 | 273950 | 271 | 1010 |
| 4 | 95950 | 103 | 931 |
| 5 | 122230 | 100 | 1222 |
| 6 | 108710 | 113 | 962 |
| 7 | 97830 | 115 | 850 |

This time, cluster 2 has the highest ratio of fast-food restaurant and population.

# Conclusions

The result of the above analyses show cluster 4 in the first analysis has the highest ratio of fast-food restaurant and population, and cluster 2 in the second analysis has the highest ratio of fast-food restaurant and population. Indeed, both cluster 4 and 2 contain the same city, Hagerstown. And the second high ratio of fast-food restaurant and population in first analysis is cluster 2 and in the second analysis is cluster 5. Both clusters contain the same cities, Gaithersburg and Germantown. Therefore, the recommendation of the first location of opening fast-food restaurant will be Hagerstown, and the second option will be Gaithersburg or Germantown.

# Future directions

We can include more data such as average income, median income, and sales in restaurant of each city to fine turn the result and predict the potential profit for the selected restaurant location. Also, we can try different methods and algorithms to compare the result of k-means clustering and see any improvement among these methods. However, the business model, local economy, and global event like COVID-19 change consumers’ behavior. These factors are also important to affect the business as well.