The Battle of Neighborhoods

Finding the best neighborhood to open a Chinese restaurant in Montgomery County of Maryland

Applied Data Science Capstone Project

Jialun Guo

1. Introduction

1.1 Background

The state of Maryland, bordering Virginia, West Virginia, and the District of Columbia, is part of the Washington D.C. metropolitan area, one of the most educated and most affluent metropolitan areas in the US. Montgomery County, is the most populous county in Maryland, which provides lots of business opportunities and business friendly environment. One business man newly relocated to Maryland. He is planning to open a Chinese restaurant in Montgomery County, where is close to his new home. This report aims to recommend some potential neighborhoods based on zip code to start his new business. This will help to give a good understanding of the business environment, strategically target the market, as well as reduce the risk.

1.2 Business Problem

The Washington D.C. metropolitan area is famous for its diversified cuisine because that most Embassies are located in the capital city and Montgomery County. So it is evident that to survive in such a highly competitive market, any new venture needs to be analyzed carefully. Various factors need to be studied in order to decide the location:

1. Montgomery County segmentation of neighborhood
2. Montgomery County Demographics
3. What is the power of consumption of the neighborhood
4. Are there any other venues like entertainment zones, shopping malls, grocery stores etc nearby where floating people is high
5. Are there any strong competitors already existing

2. Data

2.1 Python packages and Dependencies

- Pandas - Library for Data Analysis
- NumPy – Library to handle data in a vectorized manner
- JSON – Library to handle JSON files
- Geopy – To retrieve Location Data
- Requests – Library to handle http requests
- Matplotlib – Python Plotting Module
- Sklearn – Python machine learning Library
- Folium – Map rendering Library

2.2 Data sets

Data1: List of Montgomery County Zip codes

	ZIPCODE
0	20886
1	20895
2	20896
3	20899
4	20902

Source:https://www.zillow.com/browse/homes/md/montgomery-county/

This data set is used to determine the scope of our analysis.

Data 2: Maryland AGI by zip code

	ZIPCODE	Total Returns	Average_AGI
0	20854	22,390	252251.0
1	20815	14,380	239128.0
2	20817	16,420	229553.0
3	20816	7,470	227698.0
4	20818	890	179174.0

Source: https://datausa.io/profile/geo/maryland#economy

This data set reflects the power of consumption of each neighborhood.

Data 3: Maryland Census 2010 Data

	FID	ZCTA5CE10	FIRST_STAT	FIRST_GEOI	FIRST_CLAS	FIRST_MTFC	FIRST_FUNC	ZCTA5N	STATE	AREALAND	AREAWATR	POP100	HU100	NHW	N
0	1	20601	24	2420601	B5	G6350	S	20601	24	115635266	387684	24156	8722	9785	11
1	2	20602	24	2420602	B5	G6350	S	20602	24	35830723	352762	24955	9736	8466	13(
2	3	20603	24	2420603	B5	G6350	S	20603	24	44239637	219356	28967	10317	9625	15(
3	4	20606	24	2420606	B5	G6350	S	20606	24	7501011	1248760	431	230	377	
4	5	20607	24	2420607	B5	G6350	S	20607	24	54357590	448221	9802	3504	2165	6

Source:
https://data.imap.maryland.gov/datasets/eb706b48117b43d482c63d02017fc3ff_1

This data set demonstrates the demographics of each neighborhood.

Data 4: Maryland Political Boundaries Json Data

	the_geom	OBJECTID_1	OBJECTID	AREA	PERIMETER	STATEZIP_2	STATEZIP_3	ZIPCODE1	ZIPCODE2	ZIPNAME	Shape_Leng	Shape
0	MULTIPOLYGON (((-79.3443788816394 39.657186418...	1	1	0	65617.799321	2	1	21531.0	2.402322e+09	Friendsville	65985.296521	1.71270
1	MULTIPOLYGON (((-77.47019284905247 39.71844165...	24	24	0	3095.687185	25	24	21719.0	2.402122e+09	Cascade	3095.687185	3.67555
2	MULTIPOLYGON (((-76.6490020078736 39.617757533...	25	25	0	41796.823787	26	25	21161.0	2.400521e+09	White Hall	41796.823787	5.88854
3	MULTIPOLYGON (((-76.5221135638148 39.539045814...	73	73	0	18663.880331	74	73	21013.0	2.402521e+09	Baldwin	18663.880331	1.10893
4	MULTIPOLYGON (((-76.48823506556019 39.42093878...	525	525	0	50123.651438	526	525	21236.0	2.400521e+09	Nottingham	50123.651438	2.28802

Source:
https://data.imap.maryland.gov/datasets/maryland-political-boundaries-zip-codes-11-digit/geoservice

This data set provides coordinates of each neighborhood for Foursquare API analysis.

2.3 Foursquare API

Foursquare API has a database of more than 105 million places. This project would use Foursquare API as its prime data gathering source. Due to request limitations, the number of venues per neighborhood parameter would reasonably be set to 1000, and the radius parameter would be set to 3000.

2.4 Preprocessing

2.3.1 Create a new empty data frame to combine the data derived from different sources
2.3.2 Fill out the new data frame one by one with the data sets mentioned above
2.3.3 Reset the index for a better view where is necessary
2.3.4 Convert data types to proper format where is necessary
2.3.5 Drop duplicate rows where is necessary
2.3.6 Drop rows with null values where is necessary
2.3.7 Data Normalization transforms values of several variables into a similar range
2.3.8 Get coordinates for each neighborhood
2.3.9 Visualize the data on map for a better understanding

3. Methodology and Results

3.1 Exploratory Analysis

After the preprocessing, we get a data frame with demographics and income level information, grouped by zip code of Montgomery County.

	ZIPCODE	POPULATION	ASIAN	PNHW	PNHB	PNHAI	PNHA	PNHNH	PNHO	PNHT	PHISP	AGI
0	20886	0.514437	0.249308	37.7	23.5	0.2	10.8	0.0	0.5	2.9	24.4	54306
1	20895	0.294516	0.077887	73.7	6.3	0.1	5.9	0.0	0.4	2.5	11.1	119578
2	20902	0.754931	0.339864	35.1	17.9	0.2	10.1	0.1	0.5	2.4	33.8	54719
3	20901	0.538395	0.188106	43.1	23.3	0.1	7.8	0.1	0.4	2.6	22.6	68044
4	20904	0.844133	0.521531	26.7	41.6	0.1	13.8	0.1	0.4	2.6	14.8	62405

When we compute pairwise correlation of columns, we can get a rough liner relationship of variables:

- Asians tend to live in groups (correlation between Asians population and whole area population is 0.622)

- Asians love to live in populous neighborhoods (correlation between Asians population and percentage of non-Hispanics Asians is 0.869)

	POPULATION	ASIAN	PNHW	PNHB	PNHAI	PNHA	PNHNH	PNHO	PNHT	PHISP	AGI
POPULATION	1.000000	0.868996	-0.485764	0.298984	-0.291498	0.346048	-0.039951	0.459193	-0.184242	0.434388	-0.165317
ASIAN	0.868996	1.000000	-0.422864	0.162453	-0.252502	0.622273	-0.140338	0.247113	-0.022575	0.241611	-0.121197
PNHW	-0.485764	-0.422864	1.000000	-0.828242	0.155106	-0.534343	-0.039483	-0.598749	-0.079190	-0.778536	0.721610
PNHB	0.298984	0.162453	-0.828242	1.000000	-0.100673	0.212510	0.071655	0.437533	0.225650	0.478008	-0.646099
PNHAI	-0.291498	-0.252502	0.155106	-0.100673	1.000000	-0.266400	-0.135364	-0.187400	0.105994	-0.037619	-0.116854
PNHA	0.346048	0.622273	-0.534343	0.212510	-0.266400	1.000000	-0.019899	0.223653	0.213567	0.131455	-0.165814
PNHNH	-0.039951	-0.140338	-0.039483	0.071655	-0.135364	-0.019899	1.000000	0.172899	0.107065	0.015471	0.037859
PNHO	0.459193	0.247113	-0.598749	0.437533	-0.187400	0.223653	0.172899	1.000000	-0.188559	0.600943	-0.258774
PNHT	-0.184242	-0.022575	-0.079190	0.225650	0.105994	0.213567	0.107065	-0.188559	1.000000	-0.275419	-0.062480
PHISP	0.434388	0.241611	-0.778536	0.478008	-0.037619	0.131455	0.015471	0.600943	-0.275419	1.000000	-0.670748
AGI	-0.165317	-0.121197	0.721610	-0.646099	-0.116854	-0.165814	0.037859	-0.258774	-0.062480	-0.670748	1.000000

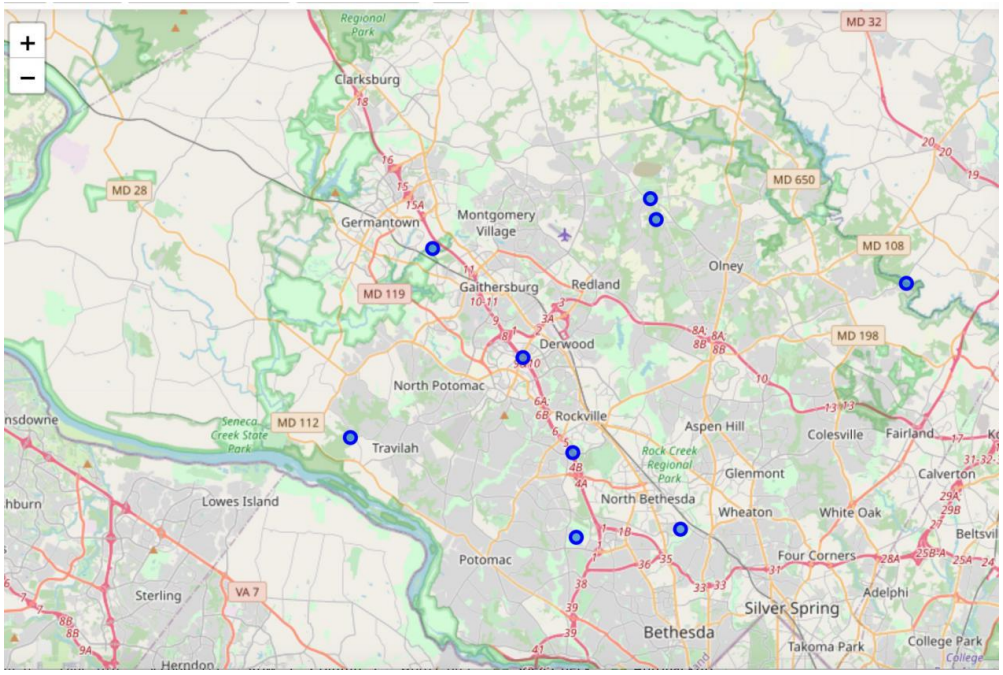
Then we pull out the top 25 neighborhoods of the most Asian population, the highest AGI, the highest percentage of Asians, and the highest population. The intersection of these four sets contains 9 neighborhoods. This will be our final data frame for next step.

3.2 Foursquare API Analysis

3.2.1 We add the coordinates for each neighborhood, and transform the data into a pandas data frame.

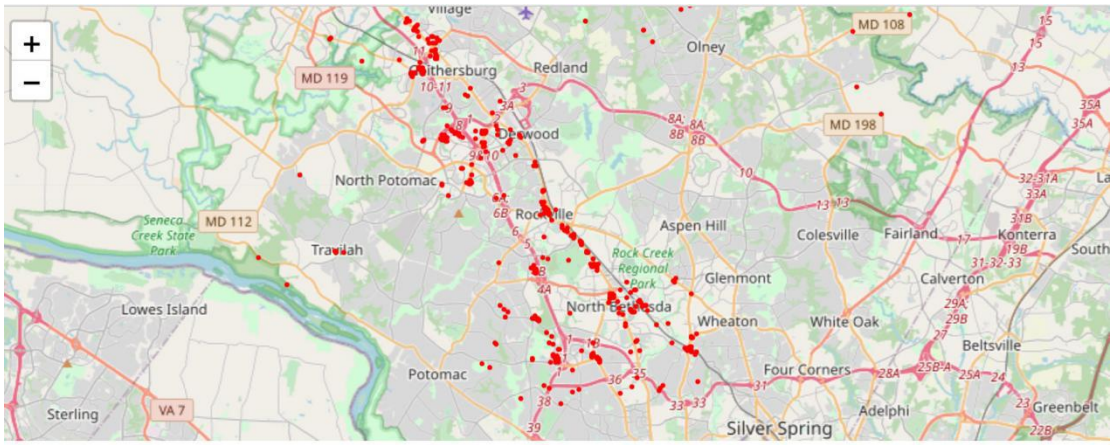
	ZIPCODE	NEIGHBORHOOD	LATITUDE	LONGITUDE
0	20855	Derwood	39.174405	-77.108489
1	20832	Olney	39.184028	-77.111894
2	20878	Gaithersburg	39.161125	-77.240420
3	20905	Silver Spring	39.145114	-76.959651
4	20850	Rockville	39.111113	-77.187244
5	20854	Potomac	39.074261	-77.289841
6	20852	Rockville	39.067175	-77.157565
7	20817	Bethesda	39.028170	-77.155508
8	20814	Bethesda	39.031823	-77.093820

Then we visualize the selected neighborhoods of Montgomery County on a map.



3.2.2 Explore Selected Neighborhoods in Montgomery County

Using the he API request URL, we get the top 1000 venues that are in the selected neighborhoods of Montgomery County with a radius of 3000 meters. By transforming the data into a python data frame, we are able to visualize all the venues on a single choropleth map.



We can find out that there are 158 unique categories from all the returned venues. Among all the categories, restaurants are the dominant ones. However, there are not many existing Chinese restaurants, which reflects a potential and promising business environment for a new Chinese restaurant.

```
Out[93]: Venue Category
         Mexican Restaurant       20
         Park                     20
         Coffee Shop              19
         American Restaurant      16
         Pizza Place              16
                                  ..
         Portuguese Restaurant     1
         Credit Union              1
         Gymnastics Gym            1
         Restaurant                1
         Ramen Restaurant          1
         Name: Venue, Length: 158, dtype: int64
```

Following is the chart of how many venues were returned for each neighborhood. Common sense tells us that the top 5 neighborhoods of most venues usually have the large flow of people, which is beneficial to a restaurant venture.

```
: ZIPCODE
  20878    100
  20852    100
  20850    100
  20817    100
  20814    100
  20832     16
  20855     15
  20854      9
  20905      4
  Name: Venue, dtype: int64
```

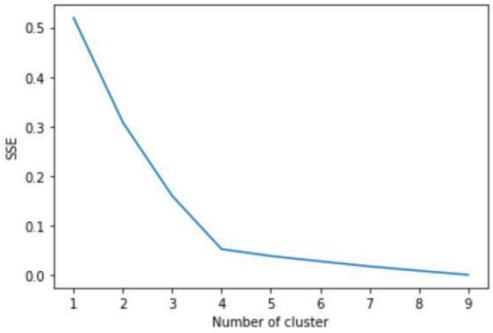
3.2.3 Analyze Each Neighborhood

We create the new data frame and display the top 10 venues for each neighborhood, to provide a brief overview of venues distribution of each neighborhood.

	ZIPCODE	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	20814	Coffee Shop	Pizza Place	Park	Chinese Restaurant	American Restaurant	Supermarket	Mexican Restaurant	Bakery	Pharmacy	Ice Cream Shop
1	20817	Cosmetics Shop	Park	Clothing Store	Coffee Shop	Mexican Restaurant	American Restaurant	Bank	Juice Bar	Boutique	Pool
2	20832	Park	Trail	Construction & Landscaping	Golf Course	Home Service	Brewery	Farm	Soccer Field	Lawyer	Theme Park
3	20850	Sandwich Place	Hotel	American Restaurant	Asian Restaurant	Mexican Restaurant	Pizza Place	Coffee Shop	Park	Italian Restaurant	Ice Cream Shop
4	20852	Mexican Restaurant	American Restaurant	Coffee Shop	Sushi Restaurant	Thai Restaurant	Peruvian Restaurant	Grocery Store	Bubble Tea Shop	Burger Joint	Chinese Restaurant
5	20854	Trail	Accessories Store	Pizza Place	Chinese Restaurant	Soccer Field	Spa	Restaurant	Grocery Store	Golf Course	Gas Station
6	20855	Park	Trail	Construction & Landscaping	Theme Park	Brewery	Photography Studio	Golf Course	Soccer Field	Outdoors & Recreation	Garden
7	20878	Sandwich Place	Ice Cream Shop	Vietnamese Restaurant	Mexican Restaurant	Donut Shop	Hotel	Pizza Place	Convenience Store	Grocery Store	Bank
8	20905	Garden Center	Basketball Court	Park	Moving Target	Yoga Studio	Filipino Restaurant	Fast Food Restaurant	Farmers Market	Farm	Event Space

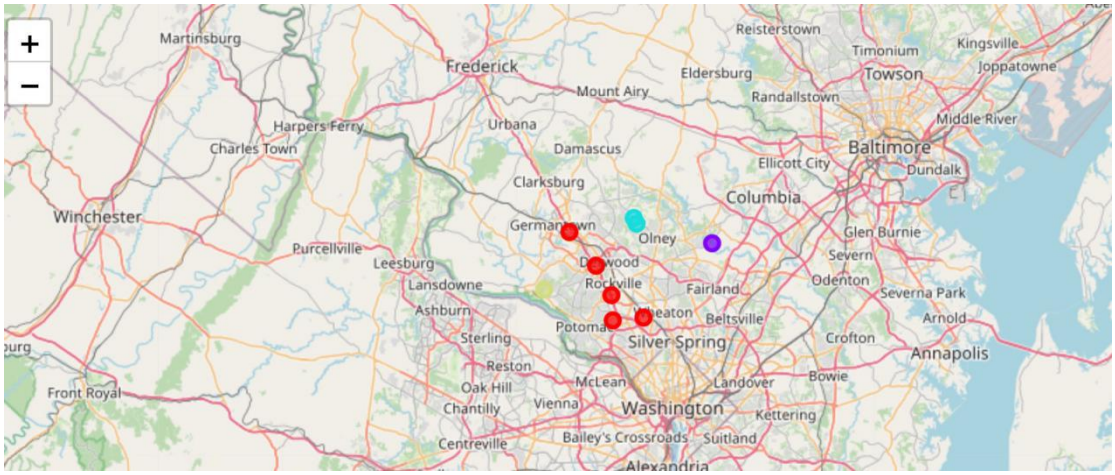
3.3 Cluster Analysis

3.3.1 Use Elbow Curve to find the optimum number of clusters, k=4.



3.3.2 Cluster Neighborhoods

After the data collection we can run k-means clustering to cluster the neighborhoods into 4 clusters and visualize all the data on a map below.



3.3.3 Examine Clusters

```
selected_montgomery_merged.loc[selected_montgomery_merged['Cluster Labels'] == 0, selected_montgomery_merged.columns[[0
```

	ZIPCODE	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	20878	Sandwich Place	Ice Cream Shop	Vietnamese Restaurant	Mexican Restaurant	Donut Shop	Hotel	Pizza Place	Convenience Store	Grocery Store	Bank
4	20850	Sandwich Place	Hotel	American Restaurant	Asian Restaurant	Mexican Restaurant	Pizza Place	Coffee Shop	Park	Italian Restaurant	Ice Cream Shop
6	20852	Mexican Restaurant	American Restaurant	Coffee Shop	Sushi Restaurant	Thai Restaurant	Peruvian Restaurant	Grocery Store	Bubble Tea Shop	Burger Joint	Chinese Restaurant
7	20817	Cosmetics Shop	Park	Clothing Store	Coffee Shop	Mexican Restaurant	American Restaurant	Bank	Juice Bar	Boutique	Pool
8	20814	Coffee Shop	Pizza Place	Park	Chinese Restaurant	American Restaurant	Supermarket	Mexican Restaurant	Bakery	Pharmacy	Ice Cream Shop

```
selected_montgomery_merged.loc[selected_montgomery_merged['Cluster Labels'] == 1, selected_montgomery_merged.columns[[0
```

	ZIPCODE	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
3	20905	Garden Center	Basketball Court	Park	Moving Target	Yoga Studio	Filipino Restaurant	Fast Food Restaurant	Farmers Market	Farm	Event Space

```
selected_montgomery_merged.loc[selected_montgomery_merged['Cluster Labels'] == 2, selected_montgomery_merged.columns[[0
```

	ZIPCODE	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	20855	Park	Trail	Construction & Landscaping	Theme Park	Brewery	Photography Studio	Golf Course	Soccer Field	Outdoors & Recreation	Garden
1	20832	Park	Trail	Construction & Landscaping	Golf Course	Home Service	Brewery	Farm	Soccer Field	Lawyer	Theme Park

```
selected_montgomery_merged.loc[selected_montgomery_merged['Cluster Labels'] == 2, selected_montgomery_merged.columns[[0
```

	ZIPCODE	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	20855	Park	Trail	Construction & Landscaping	Theme Park	Brewery	Photography Studio	Golf Course	Soccer Field	Outdoors & Recreation	Garden
1	20832	Park	Trail	Construction & Landscaping	Golf Course	Home Service	Brewery	Farm	Soccer Field	Lawyer	Theme Park

```
selected_montgomery_merged.loc[selected_montgomery_merged['Cluster Labels'] == 3, selected_montgomery_merged.columns[[0
```

	ZIPCODE	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
5	20854	Trail	Accessories Store	Pizza Place	Chinese Restaurant	Soccer Field	Spa	Restaurant	Grocery Store	Golf Course	Gas Station

4. Conclusion

4.1 Recommendation

This project recommend options with facts and data for people who plan to open an Chinese restaurant in Montgomery County of Maryland. As the business problem mentioned before, out key criteria of location decision will be the demographics, power of consumption, competitor profile, and flow of people. Given all the analysis above, we can draw a conclusion that the cluster 0, which contains 5 neighborhoods, is best for our client to open an

Chinese restaurant. Here are the zip codes of these neighborhoods: 20878, 20850, 20852, 20814, 20817.

4.2 Limitation

Location is the most important concern when opening a new restaurant. There are some other factors we need to take into consideration: safety, parking, transportation, tax policy, etc. We may discuss them later when we get other new data.