Capstone Project - The Battle of Neighborhoods

Introduction/Business Problem

In 1945, Taipei was designated as a provincial municipality. Four years later, the Chinese Nationalist Government relocated to Taiwan, and Taipei became a provisional capital. From then on, its status grew more and more important. In July, 1967, Taipei became a directly-controlled municipality. Considering the city's urban development planning, Nangang Township, Jingmei Township, Muzha Township and Neihu Township of Taipei County, along with Beitou Township and Shilin Township - managed by Yangmingshan Administration Bureau - were annexed into Taipei City a year later. Meanwhile, a plan to transform Taipei into a city of 2.5 million took shape. Population grew quickly upon Taipei's status upgrade. The city's development also started to shift eastward, and the Xinyi urban center project was formulated as a result. In 1990, Taipei's administrative districts went through another reorganization: the 16 districts were restructured into 12. They are: Songshan, Xinyi, Daan, Zhongshan, Zhongzheng, Datong, Wanhua, Wenshan, Nangang, Neihu, Shilin, and Beitou.

As the largest metropolitan area in Taiwan, Taipei city itself has the population of 2.64 million (excluding two satellite cities: New Taipei city and Keelung City) to support its booming business activities. In general, Taipei city is now one of the most exotic Asian cities in many aspects and is famous for its rich and multicultural dining selection.

A catering startup team, which is going to open a restaurant in the city and target family customers with children aging from 3-9 years old, is consulting our data team to see if we can utilize geospatial and other open data to figure out the best location to run the business. Therefore, the **business problem** that we are trying to solve is <u>to find out locations with high catering business potential for the team</u>.

The outcome of analysis should include a list of suggested neighborhoods extracted from 12 boroughs and 456 neighborhoods of Taipei city. Therefore the startup team can base on the neighborhoods to decide which one is the most suitable location to run the business.

Data description

To conduct the data analysis, the data team needs to use the following data and data services:

1. 2021 February Taipei city boroughs, neighborhoods and population data.
2. 2017 Taipei city boroughs and neighborhoods income tax data.
3. Latitude and longitude data of Taipei city boroughs and neighborhoods through Geopy library.
4. Venue data of boroughs and neighborhoods through Foursquare API.
5. Folium library to draw the map and mark the locations.

The population can be accessed at: https://data.gov.tw/dataset/136896
The annual income data can be accessed at:
https://data.gov.tw/dataset/17983

Please be noted that due to government statistics processing and financial regulation, the 2017 Taipei city boroughs and neighborhoods income tax data is the most updated one for public usage.

Methodology

- ## Data Preparation

Our business problem to be solved is how we can find out high business potential locations in the city to start up catering business through data analytic processes. Since our customer is going to open a restaurant and target young parents with kids, we expect that the population of young parents and children will be the key factors to be considered. Also, income level directly influences customers' purchasing power so we need to put this figure into our consideration as well. In general, we want to see the relationship between (1) young parents population, (2) children population and (3) income level and restaurant amounts in each neighborhood around the city, and then to predict the better locations to run the new restaurant.

So first of all, the first part of the main data set is extracted from 2021 February Taipei city boroughs, neighborhoods and population data. This data contains each neighborhood's aged population data from 0 to 100 years old. Since the new restaurant is targeting 2 customer segments (young parents with kids), we aggregate data from 0 to 12 as kid population data and 25 to 40 as parent population data, and ignore the rest.

	ID	Neighborhood	Age-0	Age-1	Age-2	Age-3	Age-4	Age-5	Age-6	Age-7	...	Age-31	Age-32	Age-33	Age-34	Age-35	Age-36	Age-37	Age-38	Age-39	Age-40
0	63000010002	莊敬里 松山區	42	47	42	45	51	54	42	41	...	63	71	68	55	62	74	79	92	101	108
1	63000010003	東榮里 松山區	33	50	43	62	78	70	80	91	...	80	80	90	68	72	99	98	110	133	120
2	63000010004	三民里 松山區	36	53	64	58	69	65	65	69	...	64	83	54	64	85	103	79	104	107	115
3	63000010005	新益里 松山區	28	36	37	26	41	39	29	33	...	56	57	50	61	67	80	86	75	83	90
4	63000010006	富錦里 松山區	35	31	48	42	50	53	80	59	...	74	66	63	58	65	77	85	83	94	82
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
451	63000120038	關渡里 北投區	77	98	96	109	129	132	124	101	...	157	184	143	142	181	221	193	234	238	239
452	63000120039	泉源里 北投區	13	11	9	18	14	14	12	16	...	33	31	30	24	28	42	41	40	36	30
453	63000120040	湖山里 北投區	8	14	8	16	5	11	9	5	...	16	16	17	31	24	24	23	26	21	24
454	63000120041	大屯里 北投區	15	9	8	11	18	18	14	13	...	15	12	9	14	25	17	29	19	32	20
455	63000120042	Hutian 北投區	6	7	7	7	7	1	5	6	...	4	16	19	11	10	16	7	17	17	7

456 rows × 43 columns

As mentioned earlier, income level is another key element to be considered. Therefore, the main data set will also merge with 2017 Taipei city boroughs and neighborhoods income tax data so that we integrate population data with annual income statistics data, including annual total income amount, income average, income median, standard deviation, etc, on the same page.

	Borough	Neighborhood	Income	IncomeAvg	IncomeMedian	1stQ	3rdQ	Std	CC	Age-0	...	Age-31	Age-32	Age-33	Age-34	Age-35	Age-36	Age-37	Age-38	A
0	萬華區	西門里 萬華區	1258566	1013	701	366	1221	1538.38	151.81	23	...	38	50	33	53	56	64	54	83	
1	萬華區	新起里 萬華區	2507677	1112	682	374	1302	2061.48	185.38	48	...	81	100	102	83	91	135	122	119	
2	萬華區	全德里 萬華區	1480791	1064	682	383	1261	1987.25	186.81	24	...	51	56	71	55	61	66	64	85	
3	萬華區	壽德里 萬華區	1521674	959	679	388	1215	978.40	102.04	38	...	58	60	61	60	64	80	75	86	
4	萬華區	萬壽里 萬華區	1171107	1070	676	353	1261	1306.62	122.06	23	...	32	34	27	33	44	51	44	47	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
446	中山區	行政里 中山區	2061119	996	630	351	1200	2056.10	206.40	53	...	93	97	92	85	109	90	115	100	
447	中山區	新庄里 中山區	1448754	880	619	355	1134	937.62	106.53	33	...	53	79	74	59	72	88	85	74	
448	中山區	正義里 中山區	1927190	1067	618	328	1086	6565.49	615.60	32	...	72	74	60	64	82	92	95	104	
449	中山區	聚盛里 中山區	1354314	954	607	343	1136	1349.87	141.43	27	...	60	44	50	54	69	67	67	81	
450	中山區	大佳里 中山區	258318	850	560	331	1039	1055.59	124.23	3	...	13	18	14	13	16	10	18	19	

451 rows × 50 columns

Second, we use the Geopy library with borough and neighborhood names to retrieve borough and neighborhood's latitude and longitude. This will be the input parameter of Foursquare API.

	Borough	Neighborhood	Latitude	Longitude	Income	IncomeAvg	IncomeMedian	1stQ	3rdQ	Std	...	Age-31	Age-32	Age-33	Age-34	Age-35	Age-36	Age-37
0	萬華區	西門里 萬華區	25.042815	121.505472	1258566	1013	701	366	1221	1538.38	...	38	50	33	53	56	64	54
1	萬華區	新起里 萬華區	25.041003	121.505049	2507677	1112	682	374	1302	2061.48	...	81	100	102	83	91	135	122
2	萬華區	全德里 萬華區	25.023361	121.498731	1480791	1064	682	383	1261	1987.25	...	51	56	71	55	61	66	64
3	萬華區	壽德里 萬華區	25.023295	121.500671	1521674	959	679	388	1215	978.40	...	58	60	61	60	64	80	75
4	萬華區	萬壽里 萬華區	25.044793	121.505332	1171107	1070	676	353	1261	1306.62	...	32	34	27	33	44	51	44
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
446	中山區	行政里 中山區	25.065368	121.534927	2061119	996	630	351	1200	2056.10	...	93	97	92	85	109	90	115
447	中山區	新庄里 中山區	25.070702	121.530619	1448754	880	619	355	1134	937.62	...	53	79	74	59	72	88	85
448	中山區	正義里 中山區	25.050614	121.526592	1927190	1067	618	328	1086	6565.49	...	72	74	60	64	82	92	95
449	中山區	聚盛里 中山區	25.059172	121.525247	1354314	954	607	343	1136	1349.87	...	60	44	50	54	69	67	67
450	中山區	大佳里 中山區	25.072798	121.542440	258318	850	560	331	1039	1055.59	...	13	18	14	13	16	10	18

451 rows × 52 columns

Since our analysis typically focuses on children and parents customer segments, therefore we aggregate population amounts between 0 and 12 years old as kid population, and also do the same aggregation to 25 to 40 year-old population as parents population.

	Borough	Neighborhood	Latitude	Longitude	Kid	Parent	Income	IncomeAvg	IncomeMedian	1stQ	...	Age-31	Age-32	Age-33	Age-34	Age-35	Age-36	Age-37
0	萬華區	西門里 萬華區	25.042815	121.505472	343	770	1258566	1013	701	366	...	38	50	33	53	56	64	54
1	萬華區	新起里 萬華區	25.041003	121.505049	555	1506	2507677	1112	682	374	...	81	100	102	83	91	135	122
2	萬華區	全德里 萬華區	25.023361	121.498731	547	987	1480791	1064	682	383	...	51	56	71	55	61	66	64
3	萬華區	壽德里 萬華區	25.023295	121.500671	531	1116	1521674	959	679	388	...	58	60	61	60	64	80	75
4	萬華區	萬壽里 萬華區	25.044793	121.505332	253	576	1171107	1070	676	353	...	32	34	27	33	44	51	44
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
446	中山區	行政里 中山區	25.065368	121.534927	471	1491	2061119	996	630	351	...	93	97	92	85	109	90	115
447	中山區	新庄里 中山區	25.070702	121.530619	431	1141	1448754	880	619	355	...	53	79	74	59	72	88	85
448	中山區	正義里 中山區	25.050614	121.526592	442	1219	1927190	1067	618	328	...	72	74	60	64	82	92	95
449	中山區	聚盛里 中山區	25.059172	121.525247	300	909	1354314	954	607	343	...	60	44	50	54	69	67	67
450	中山區	大佳里 中山區	25.072798	121.542440	132	227	258318	850	560	331	...	13	18	14	13	16	10	18

451 rows × 54 columns

We also use the Foursquare "**search**" API endpoint by inputting latitude, longitude, "Food" category ID and other attributes to search relevant venue data of each neighborhood.

Food
4d4b7105d754a06374d81259

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	西門里 萬華區	25.042815	121.505472	SOL bistro (SOL bistro 料理小酒館)	25.041925	121.506239	Bistro
1	西門里 萬華區	25.042815	121.505472	阿財虱目魚肚	25.041700	121.505225	Taiwanese Restaurant
2	西門里 萬華區	25.042815	121.505472	Starbucks Coffee (星巴克)	25.043141	121.504920	Coffee Shop
3	西門里 萬華區	25.042815	121.505472	McDonald's (麥當勞)	25.044827	121.505363	Fast Food Restaurant
4	西門里 萬華區	25.042815	121.505472	繼光香香雞 Ji Guang Fried Chicken	25.042785	121.507610	Fried Chicken Joint
...	...	...	...	...	...	...	...
19377	大佳里 中山區	25.072798	121.542440	Mary Jane's Pizza	25.074130	121.538730	Italian Restaurant
19378	大佳里 中山區	25.072798	121.542440	松山機場 P.r.o. Coffee	25.072244	121.543275	Café
19379	大佳里 中山區	25.072798	121.542440	Ln. 180	25.072117	121.538444	Café
19380	大佳里 中山區	25.072798	121.542440	美而美漢堡三明治	25.068148	121.539484	Breakfast Spot
19381	大佳里 中山區	25.072798	121.542440	牛之鬼牛排館	25.071118	121.538388	American Restaurant

19382 rows × 7 columns

- Data Cluster

Once the catering venue data is ready, we group, summarize and count venues as restaurant density for each neighborhood and cluster all neighborhoods into **50 clusters** with restaurant density, income data and segmented population data.

- Data Description

By averaging and describing each cluster's statistics, we have observed that there is a positive correlation between restaurant density and other independent variables, especially (1) parent segment and (2) neighborhood income median. As a result we are going to establish a multiple regression model with testing data to predict each cluster's restaurant density to see how the model fits.

	Restaurant Density	Kid	Parent	Income	IncomeAvg	IncomeMedian	1stQ	3rdQ	Std	CC
Restaurant Density	1.000000	0.433183	0.545368	0.461878	0.302413	0.562527	0.550701	0.561594	0.160976	0.175756
Kid	0.433183	1.000000	0.873243	0.739018	0.197086	0.850324	0.867898	0.830686	-0.022002	0.089589
Parent	0.545368	0.873243	1.000000	0.533767	-0.037351	0.679080	0.725282	0.663128	-0.203591	-0.124297
Income	0.461878	0.739018	0.533767	1.000000	0.734548	0.883957	0.851989	0.902971	0.527937	0.617405
IncomeAvg	0.302413	0.197086	-0.037351	0.734548	1.000000	0.524160	0.474301	0.542043	0.878655	0.888699
IncomeMedian	0.562527	0.850324	0.679080	0.883957	0.524160	1.000000	0.968333	0.993011	0.250445	0.354115
1stQ	0.550701	0.867898	0.725282	0.851989	0.474301	0.968333	1.000000	0.950124	0.245581	0.335019
3rdQ	0.561594	0.830686	0.663128	0.902971	0.542043	0.993011	0.950124	1.000000	0.272993	0.374682
Std	0.160976	-0.022002	-0.203591	0.527937	0.878655	0.250445	0.245581	0.272993	1.000000	0.971924
CC	0.175756	0.089589	-0.124297	0.617405	0.888699	0.354115	0.335019	0.374682	0.971924	1.000000

- Data Modeling and Prediction
    - Data Modeling

As described above, a multiple regression model is to be established for our analysis. The target data set is the clustered neighborhood data which is aggregated into 50 groups. Then we split into train data set and test data set (ratio=0.3) and specifically adopt polynomial features to independant variables with Ridge regression to create the model.

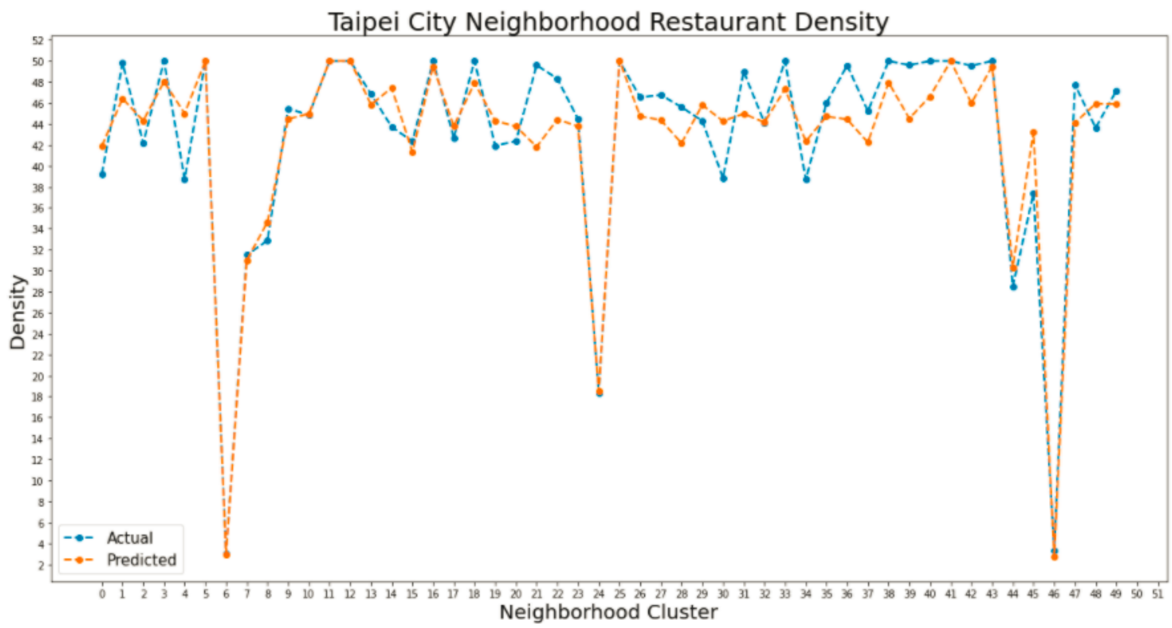
    - Data Prediction

As described above, we predict the expected restaurant density (average amount of catering business) for each neighborhood cluster. The prediction quality and result are shown as following:

```
X size: 50
y size: 50
X_train size: 35
y_train size: 35
X_test size: 15
y_test size: 15
Mean squared error: 11.63
Mean absolute error: 2.81
Coefficient of determination: 0.91
```

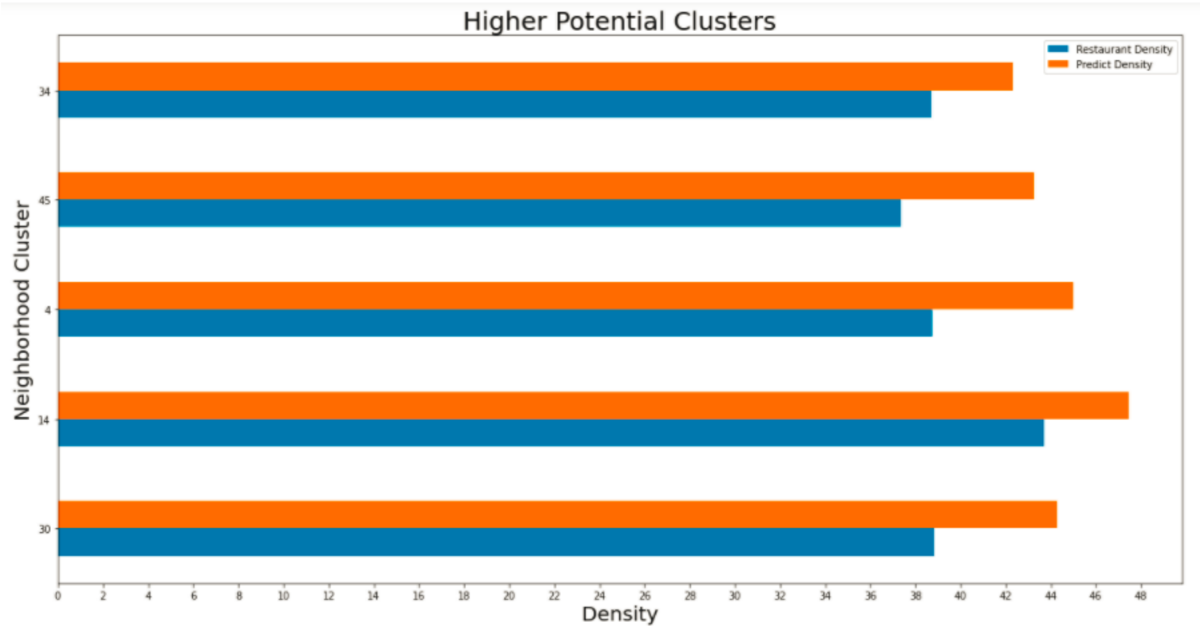
Results

The overall predicted result versus actual data is illustrated as below. We also notice some clusters' actual average restaurant density is less than the predicted value:



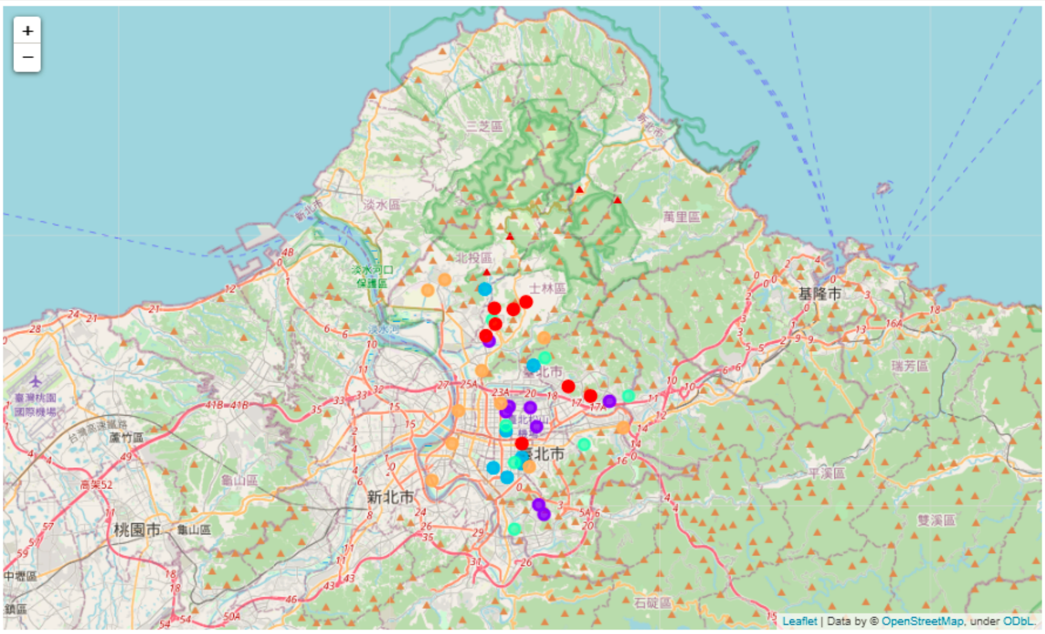
And we can further focus on those clusters because we think these clusters of neighborhoods may be regarded as “higher growth potential” in catering business.

To narrow down our analysis scope, among these clusters we pay attention to those with differences greater than 3. We recommend these clusters could be the good candidates for the startup team to investigate business opportunities since we believe there should be higher growth rates inside these clusters to run a new catering business.



We also put these neighborhoods on the city map to illustrate cluster and geographical information.

Out[296]:



Discussion

There are some points to be discussed:

1. As required we use Foursquare API to search “Food” category venue data for our analysis. It could be a good alternative if we swap to Google Map API since it may provide more local data.
2. In our model we use two types of data as our independent variables, which are population (in parent and child age) and income data. There are other types of data, such as marital status, education level, rent level,etc to be leveraged so that we can further improve the prediction accuracy.
3. We have also extracted the venue types in our dataset. Therefore, as the next phase we can extend our analysis to reveal the good locations to fit a particular type of catering business.

Conclusion

Location is the most important consideration to start a new business. We have tried to use government open data with geospatial services to conduct a simple but intuitive data analysis to provide a list of potential locations to run a new business. We believe this outcome is a good foundation to assist our customer to define their go-to-market plan more efficiently.