Opportunity to Opening New Restaurant Business on West Java

IBM DATA SCIENCE CAPSTONE FINAL PROJECT

BY MUHAMMAD HIKMATYAR

Introduction

West Java as the most populous province and **one of the biggest economy in Indonesia** has many potential to start a new business, especially a **culinary business.** There are a lot of restaurants that scattered across the province, many of them is a local restaurant like indonesian or sundanese restaurant. Although there are also a lot of Chinese Restaurant, but this type of restaurant still less a lot compared to the local restaurant. This project aims to help people, particularly to a new started businessman to open their new Chinese Restaurant on West Java. It will help them to make their business decision easily based on the **distribution of the Chinese Restaurant** on West Java. In this project, I'm creating a **hyphotetical assumptions** that doesn't include the other variables to consider such as economy outlook of the city (inflation rate, unemployment number, etc) or the market behavior of people in the city which are also an important consideration. Nonetheless, this recomendation is still an important consideration to make this business decision.

Data Section

The dataset is from https://simplemaps.com/data/id-cities. The dataset consist of 8,912 prominent cities in Indonesia, their province, and also the longitude and latitude of each cities and other relevant information. Then the data is filtered to find only the city on West Java (or 'Jawa Barat' on the dataset). The final data that will be used in this project would be the list of the cities in West Java and their longitude and latitude. Because the limitation of API calls we use (will be explained later), the city list is intentionally limited to only the top 300 cities on West Java based on the population. This is the map of the West Java and its top 300 city (representing as red circle)

Foursquare API

Foursquare API is used to get the nearby Chinese Restaurant using its 'explore' API calls. This is one of the most powerful location services API. This Foursquare API allows us to find all venues and events within an area of interest, including Chinese Restaurant as long as the geospatial information such as longitude and latitude is provided.

Data Section

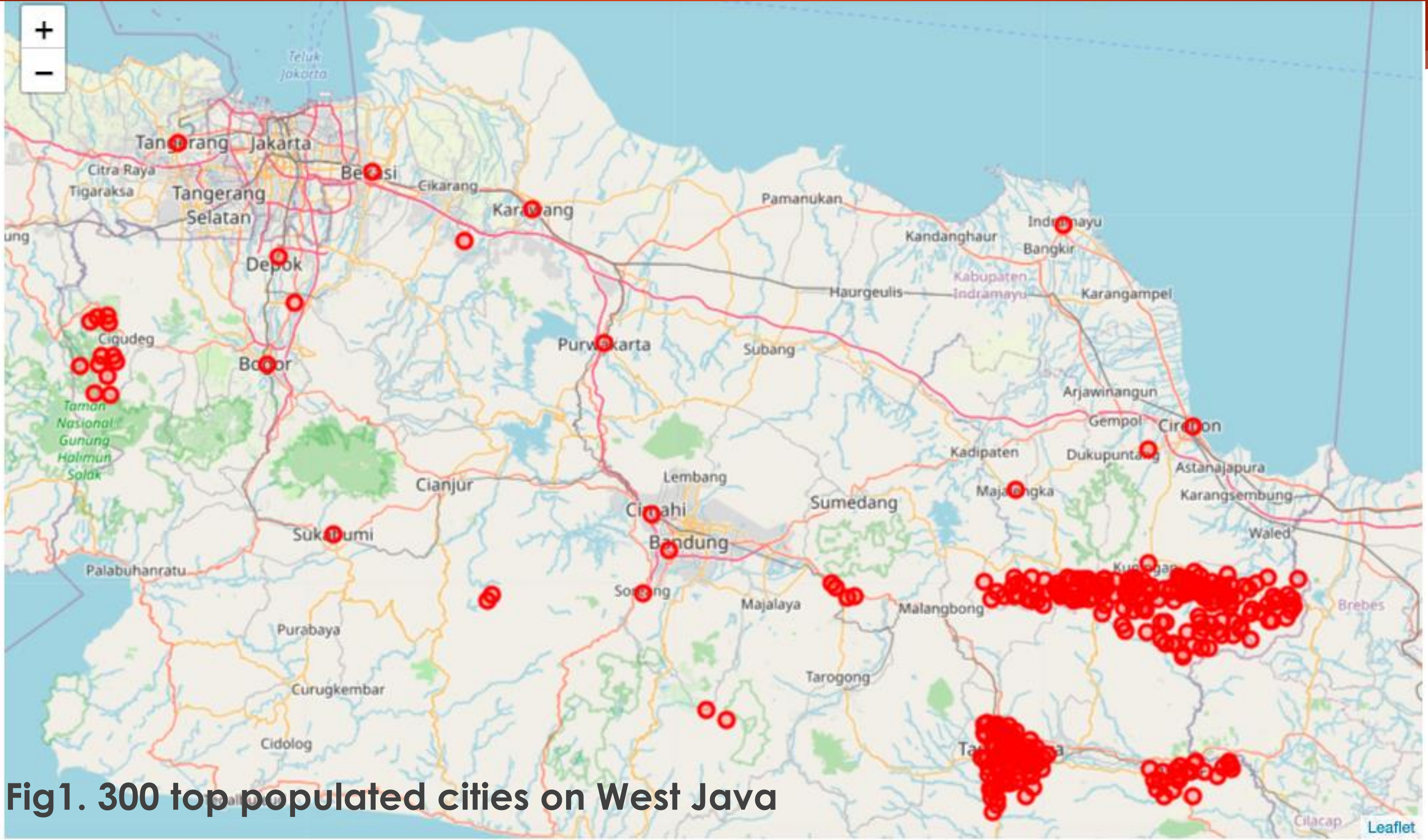


Fig1. 300 top populated cities on West Java

Methodology

Tools that used in this project:

- Indonesia Cities Geospatial data (https://simplemaps.com/data/id-cities)
- Foursquare API
- Folium Map
- Kmeans Clustering

The aim is to find the Chinese Restaurant venues around each city. The Foursquare API is used to do that with 'explore' API call to get the venues around an area of interest, as long as the longitude and latitude information is provided. The longitude and latitude information is got from the Indonesia Cities Geospatial dataset that had to be downloaded as csv file from https://simplemaps.com/data/id-cities website. After that, data is clustered using KMeans method based on the amount of Chinese Restaurant for each city. Finally the Folium Map is used visualize the clustering result on actual map.

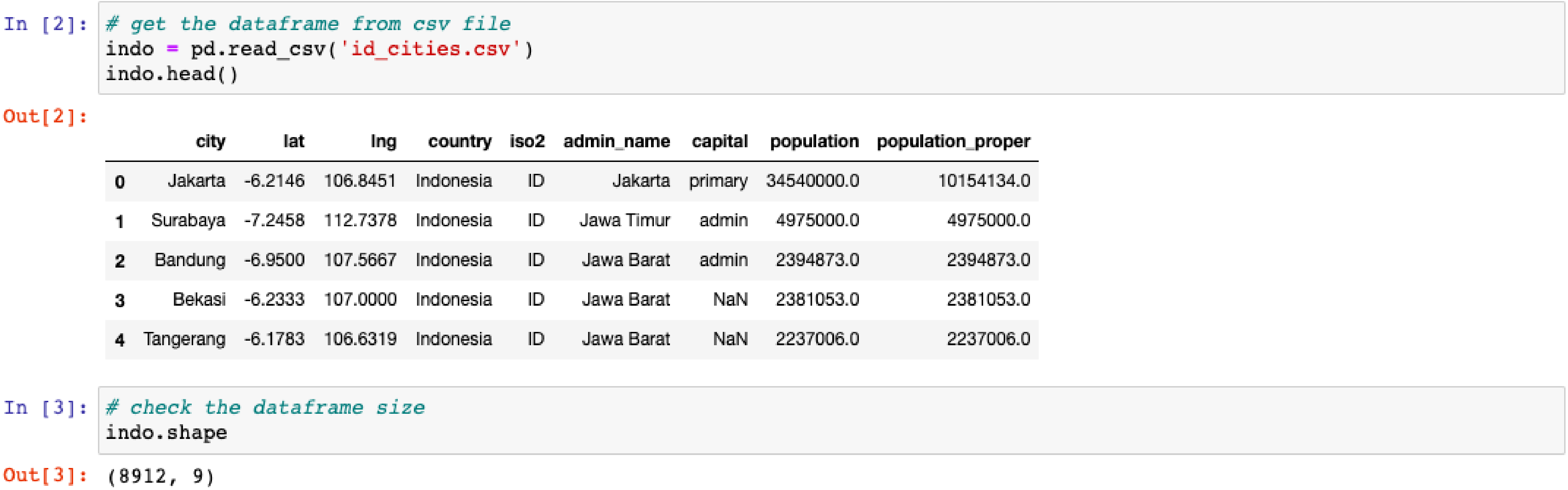
Methodology

Data Collection

In this stage, the data is collected from csv file 'id_cities.csv' that had been downloaded from https://simplemaps.com/data/id-cities website. This data consist of 8,912 prominent cities in Indonesia, their province, and also the longitude and latitude of each cities and other relevant information. Pandas library is used to read the file using the pd.read_csv() method.

Methodology

Data Collection



```
In [2]: # get the dataframe from csv file
        indo = pd.read_csv('id_cities.csv')
        indo.head()
```

Out[2]:

	city	lat	lng	country	iso2	admin_name	capital	population	population_proper
0	Jakarta	-6.2146	106.8451	Indonesia	ID	Jakarta	primary	34540000.0	10154134.0
1	Surabaya	-7.2458	112.7378	Indonesia	ID	Jawa Timur	admin	4975000.0	4975000.0
2	Bandung	-6.9500	107.5667	Indonesia	ID	Jawa Barat	admin	2394873.0	2394873.0
3	Bekasi	-6.2333	107.0000	Indonesia	ID	Jawa Barat	NaN	2381053.0	2381053.0
4	Tangerang	-6.1783	106.6319	Indonesia	ID	Jawa Barat	NaN	2237006.0	2237006.0

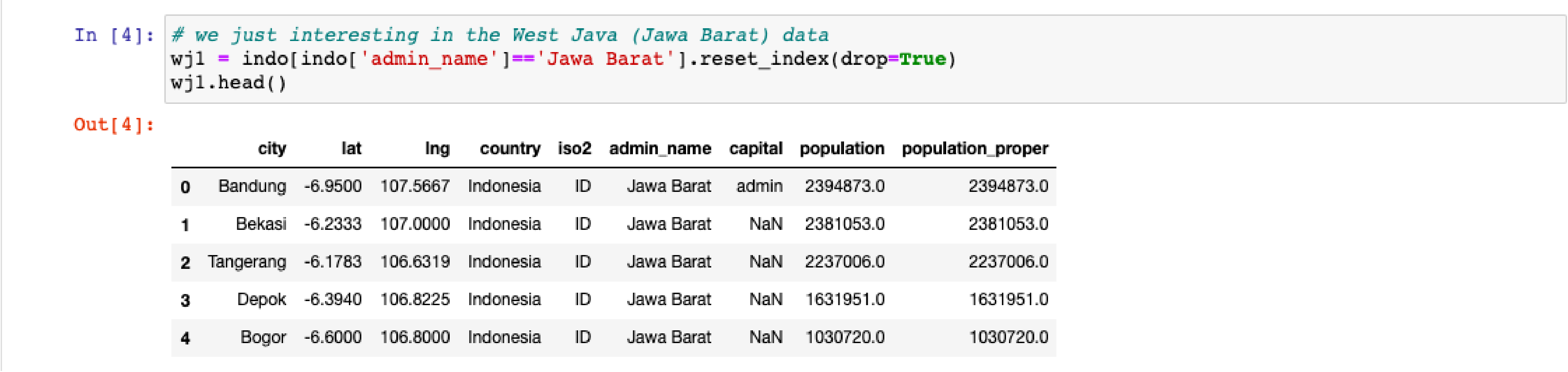
```
In [3]: # check the dataframe size
        indo.shape

Out[3]: (8912, 9)
```

Methodology

Data Preprocessing

There are 8912 data from the dataset. Since the area of interest is on West Java (or 'Jawa Barat'), the other data will be dropped.



In [4]:

```
# we just interesting in the West Java (Jawa Barat) data
wjl = indo[indo['admin_name']=='Jawa Barat'].reset_index(drop=True)
wjl.head()
```

Out[4]:

	city	lat	lng	country	iso2	admin_name	capital	population	population_proper
0	Bandung	-6.9500	107.5667	Indonesia	ID	Jawa Barat	admin	2394873.0	2394873.0
1	Bekasi	-6.2333	107.0000	Indonesia	ID	Jawa Barat	NaN	2381053.0	2381053.0
2	Tangerang	-6.1783	106.6319	Indonesia	ID	Jawa Barat	NaN	2237006.0	2237006.0
3	Depok	-6.3940	106.8225	Indonesia	ID	Jawa Barat	NaN	1631951.0	1631951.0
4	Bogor	-6.6000	106.8000	Indonesia	ID	Jawa Barat	NaN	1030720.0	1030720.0

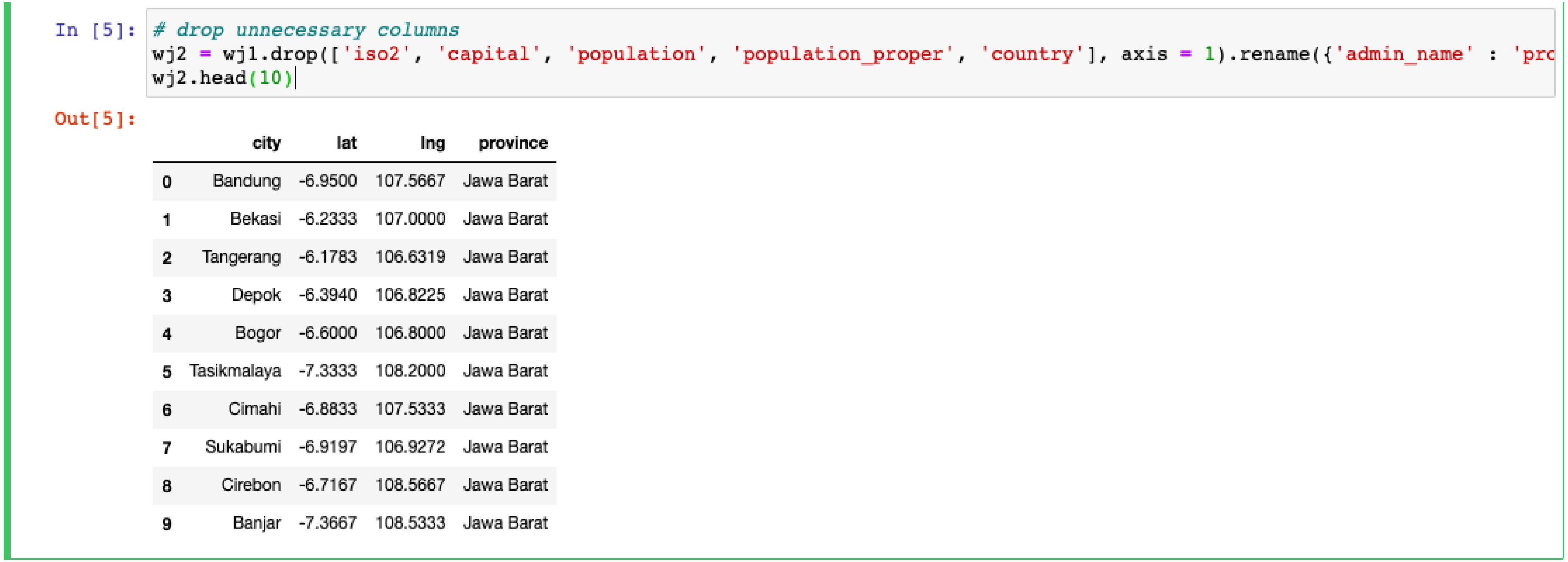
Methodology

Feature Selection

Only 'city', 'lat', 'lng', and 'admin_name' features are needed. The other features can be dropped using drop() method from Pandas library. Also the 'admin_name' feature can be renamed to 'province' using rename() method that also from Pandas library.

Methodology

Feature Selection



```
In [5]: # drop unnecessary columns
        wj2 = wj1.drop(['iso2', 'capital', 'population', 'population_proper', 'country'], axis = 1).rename({'admin_name' : 'pro
        wj2.head(10)
```

Out[5]:

	city	lat	lng	province
0	Bandung	-6.9500	107.5667	Jawa Barat
1	Bekasi	-6.2333	107.0000	Jawa Barat
2	Tangerang	-6.1783	106.6319	Jawa Barat
3	Depok	-6.3940	106.8225	Jawa Barat
4	Bogor	-6.6000	106.8000	Jawa Barat
5	Tasikmalaya	-7.3333	108.2000	Jawa Barat
6	Cimahi	-6.8833	107.5333	Jawa Barat
7	Sukabumi	-6.9197	106.9272	Jawa Barat
8	Cirebon	-6.7167	108.5667	Jawa Barat
9	Banjar	-7.3667	108.5333	Jawa Barat

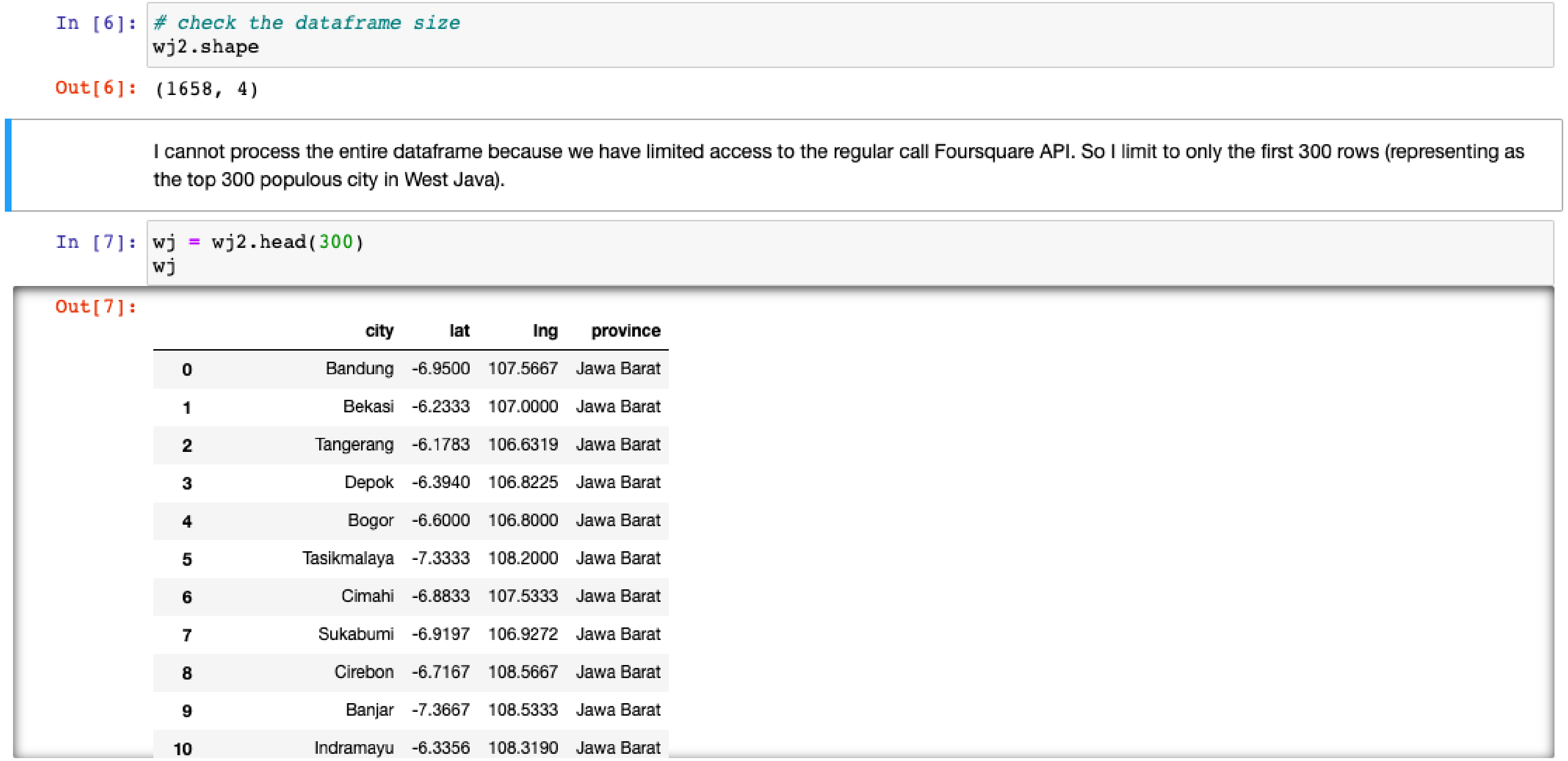
Methodology

Feature Engineering

There are 1658 city in West Java. Because of the limitation of Forsquare API calls (only 950 regular calls per day), the city list is intentionally limited to only the top 300 cities on West Java based on the population.

Methodology

Feature Engineering



```
In [6]: # check the dataframe size
        wj2.shape
Out[6]: (1658, 4)
```

I cannot process the entire dataframe because we have limited access to the regular call Foursquare API. So I limit to only the first 300 rows (representing as the top 300 populous city in West Java).

```
In [7]: wj = wj2.head(300)
        wj
Out[7]:
```

	city	lat	lng	province
0	Bandung	-6.9500	107.5667	Jawa Barat
1	Bekasi	-6.2333	107.0000	Jawa Barat
2	Tangerang	-6.1783	106.6319	Jawa Barat
3	Depok	-6.3940	106.8225	Jawa Barat
4	Bogor	-6.6000	106.8000	Jawa Barat
5	Tasikmalaya	-7.3333	108.2000	Jawa Barat
6	Cimahi	-6.8833	107.5333	Jawa Barat
7	Sukabumi	-6.9197	106.9272	Jawa Barat
8	Cirebon	-6.7167	108.5667	Jawa Barat
9	Banjar	-7.3667	108.5333	Jawa Barat
10	Indramayu	-6.3356	108.3190	Jawa Barat

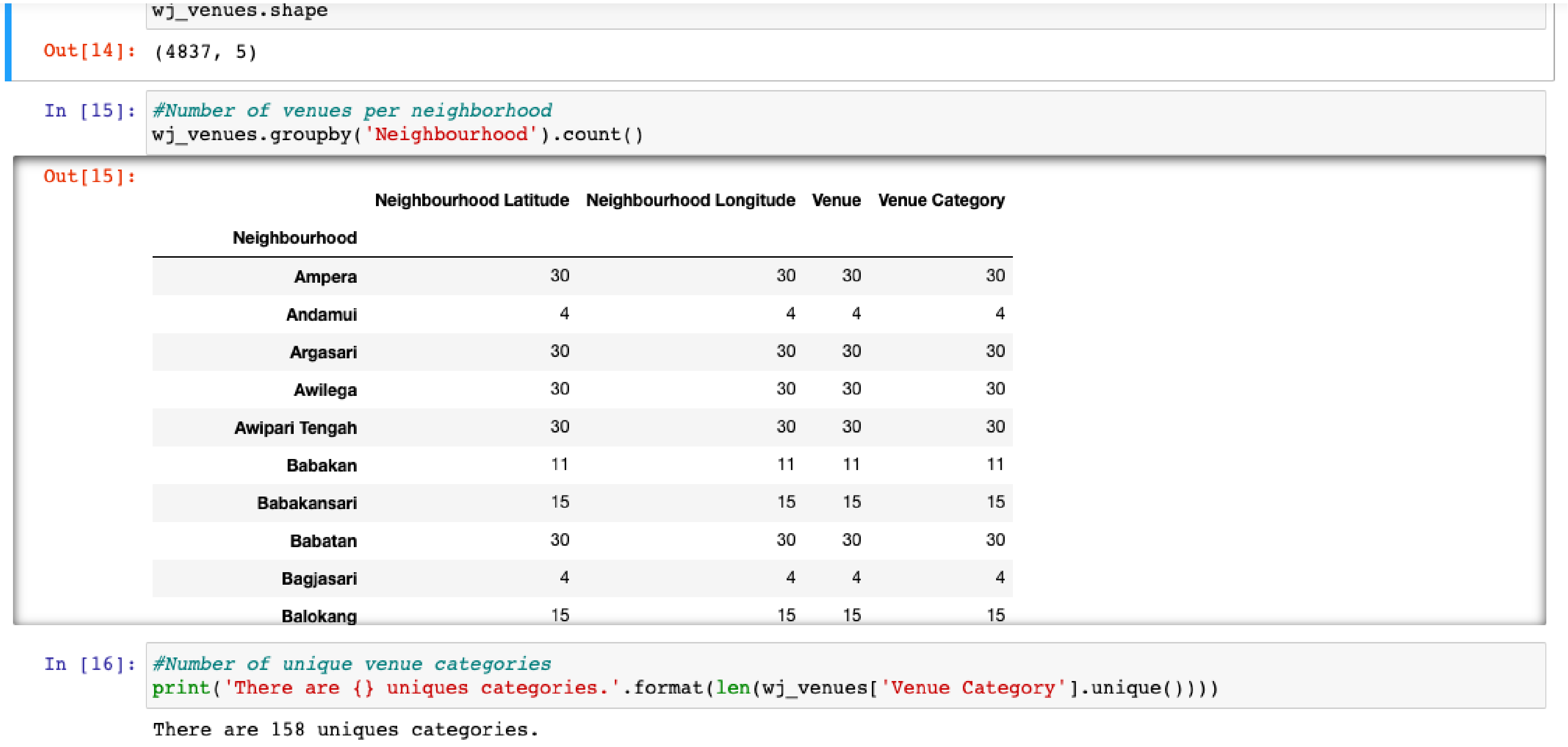
Methodology

Foursquare API Calls

The 'explore' call is used to find all venues and events within an area of interest, including Chinese Restaurant. A function is made to process every city on the list. After that, the 'Venue Category' is obtained and can be counted for each of city.

Methodology

Foursquare API Calls



```
wj_venues.shape
```

```
Out[14]: (4837, 5)
```

```
In [15]: #Number of venues per neighborhood
         wj_venues.groupby('Neighbourhood').count()
```

Out[15]:

	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Category
Neighbourhood				
Ampera	30	30	30	30
Andamui	4	4	4	4
Argasari	30	30	30	30
Awilega	30	30	30	30
Awipari Tengah	30	30	30	30
Babakan	11	11	11	11
Babakansari	15	15	15	15
Babatan	30	30	30	30
Bagjasari	4	4	4	4
Balokang	15	15	15	15

```
In [16]: #Number of unique venue categories
         print('There are {} uniques categories.'.format(len(wj_venues['Venue Category'].unique())))

         There are 158 uniques categories.
```

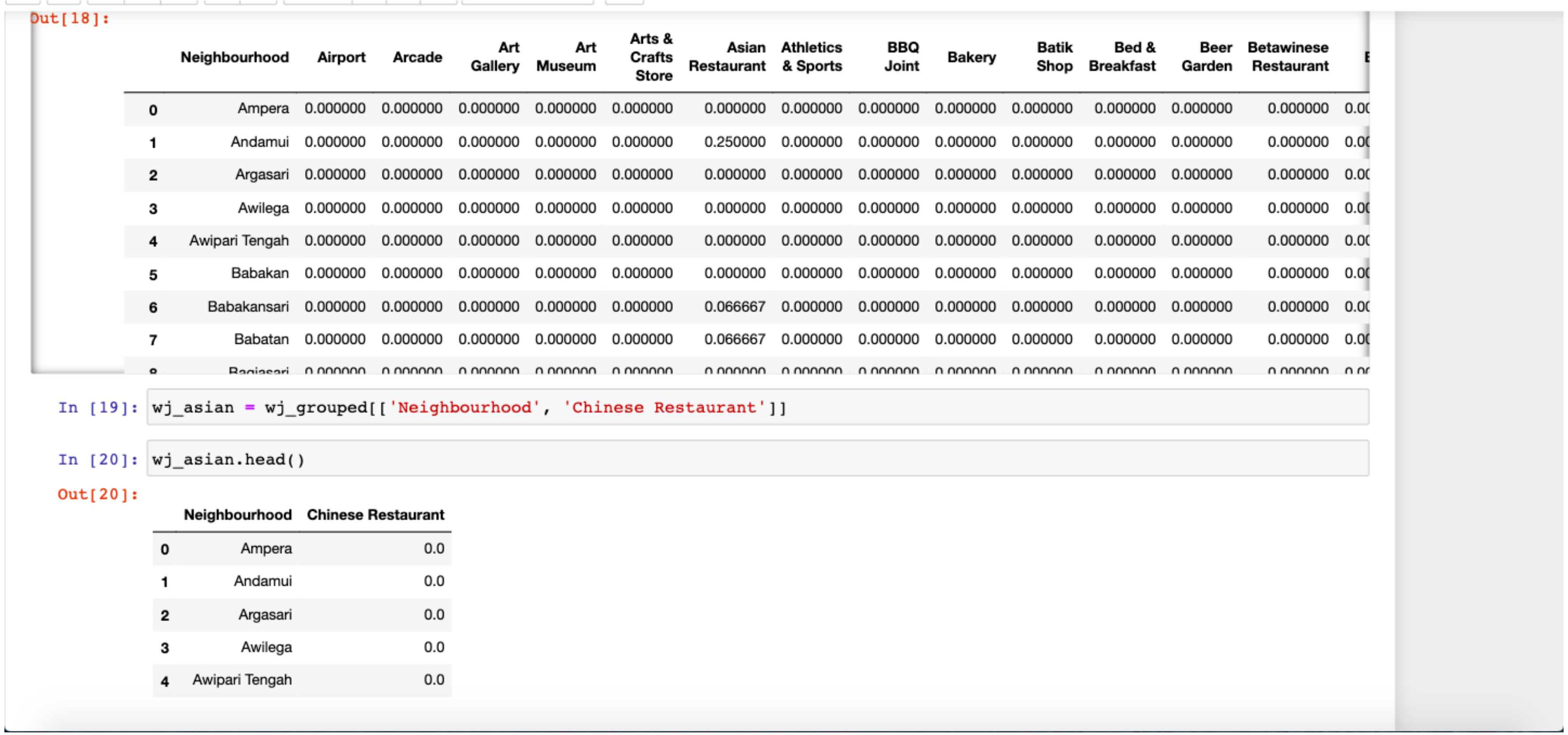
Methodology

One Hot Encoding

One hot encoding is a method to convert the categorical data into numeric data. In this project, one hot encoding is used to calculate the weight of each 'Venues Category' to each city, representing how much the certain Venues Category is appeared on each city. After converting each 'Venue Category' to numerical values using get_dummies method on pandas, pandas mean method can be used to find this values. And then after that, the dataframe can be filtered to only the 'Chinese Restaurant' column, because that's the interest of this project.

Methodology

One Hot Encoding



Out[18]:

	Neighbourhood	Airport	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	BBQ Joint	Bakery	Batik Shop	Bed & Breakfast	Beer Garden	Betawinese Restaurant	E
0	Ampera	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00
1	Andamui	0.000000	0.000000	0.000000	0.000000	0.000000	0.250000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0(
2	Argasari	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0(
3	Awilega	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0(
4	Awipari Tengah	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0(
5	Babakan	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0(
6	Babakansari	0.000000	0.000000	0.000000	0.000000	0.000000	0.066667	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0(
7	Babatan	0.000000	0.000000	0.000000	0.000000	0.000000	0.066667	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0(
8	Bagiasari	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0(

```
In [19]: wj_asian = wj_grouped[['Neighbourhood', 'Chinese Restaurant']]
```

```
In [20]: wj_asian.head()
```

Out[20]:

	Neighbourhood	Chinese Restaurant
0	Ampera	0.0
1	Andamui	0.0
2	Argasari	0.0
3	Awilega	0.0
4	Awipari Tengah	0.0

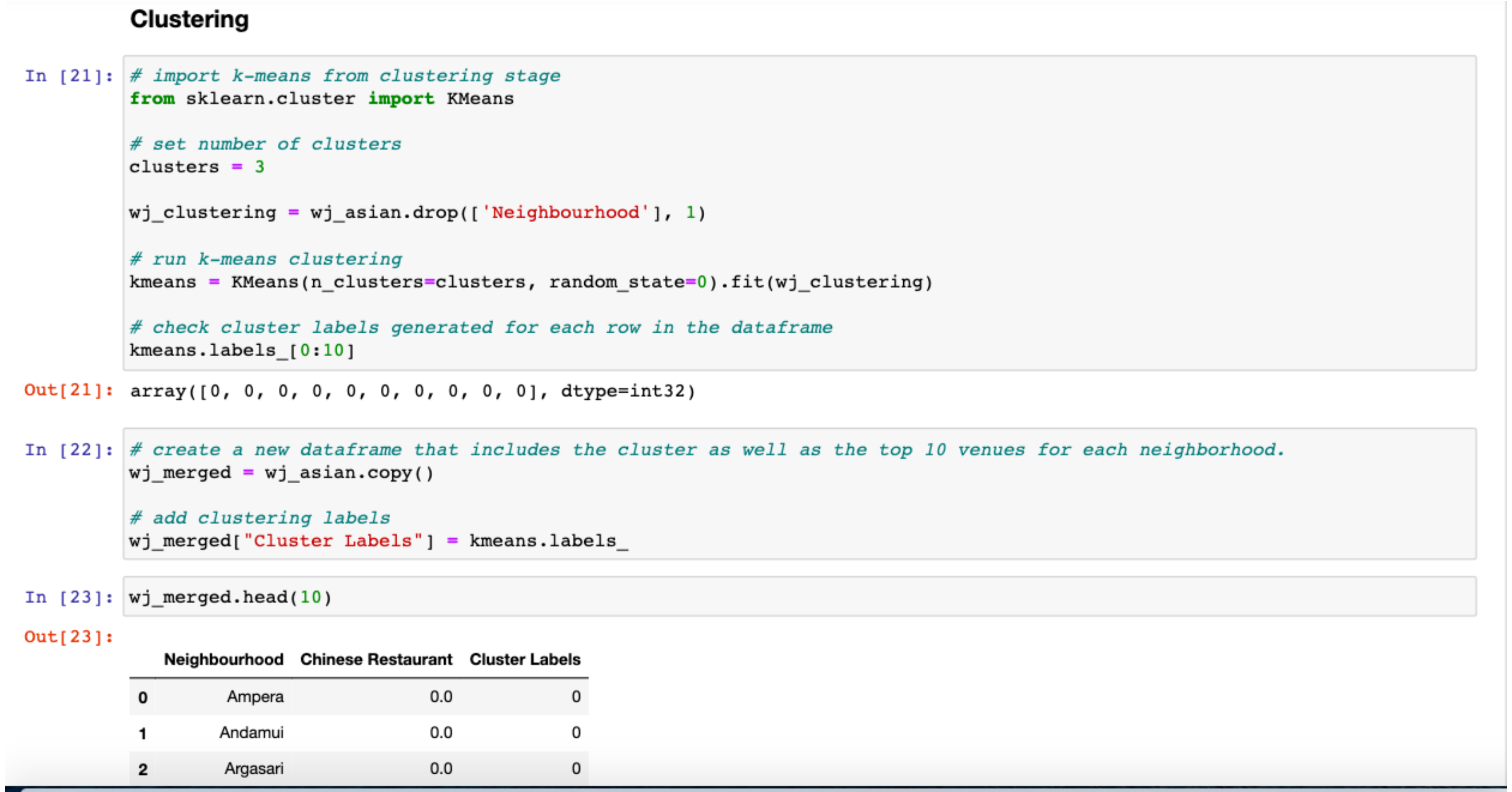
Methodology

KMeans Clustering

This is the most important part since we aim to cluster the city based on the Chinese Restaurant distribution. We can cluster them into 3 cluster (also to see its pattern).

Methodology

KMeans Clustering



Clustering

```
In [21]: # import k-means from clustering stage
         from sklearn.cluster import KMeans

         # set number of clusters
         clusters = 3

         wj_clustering = wj_asian.drop(['Neighbourhood'], 1)

         # run k-means clustering
         kmeans = KMeans(n_clusters=clusters, random_state=0).fit(wj_clustering)

         # check cluster labels generated for each row in the dataframe
         kmeans.labels_[0:10]

Out[21]: array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0], dtype=int32)
```

```
In [22]: # create a new dataframe that includes the cluster as well as the top 10 venues for each neighborhood.
         wj_merged = wj_asian.copy()

         # add clustering labels
         wj_merged["Cluster Labels"] = kmeans.labels_
```

```
In [23]: wj_merged.head(10)
```

Out[23]:

	Neighbourhood	Chinese Restaurant	Cluster Labels
0	Ampera	0.0	0
1	Andamui	0.0	0
2	Argasari	0.0	0

RESULT

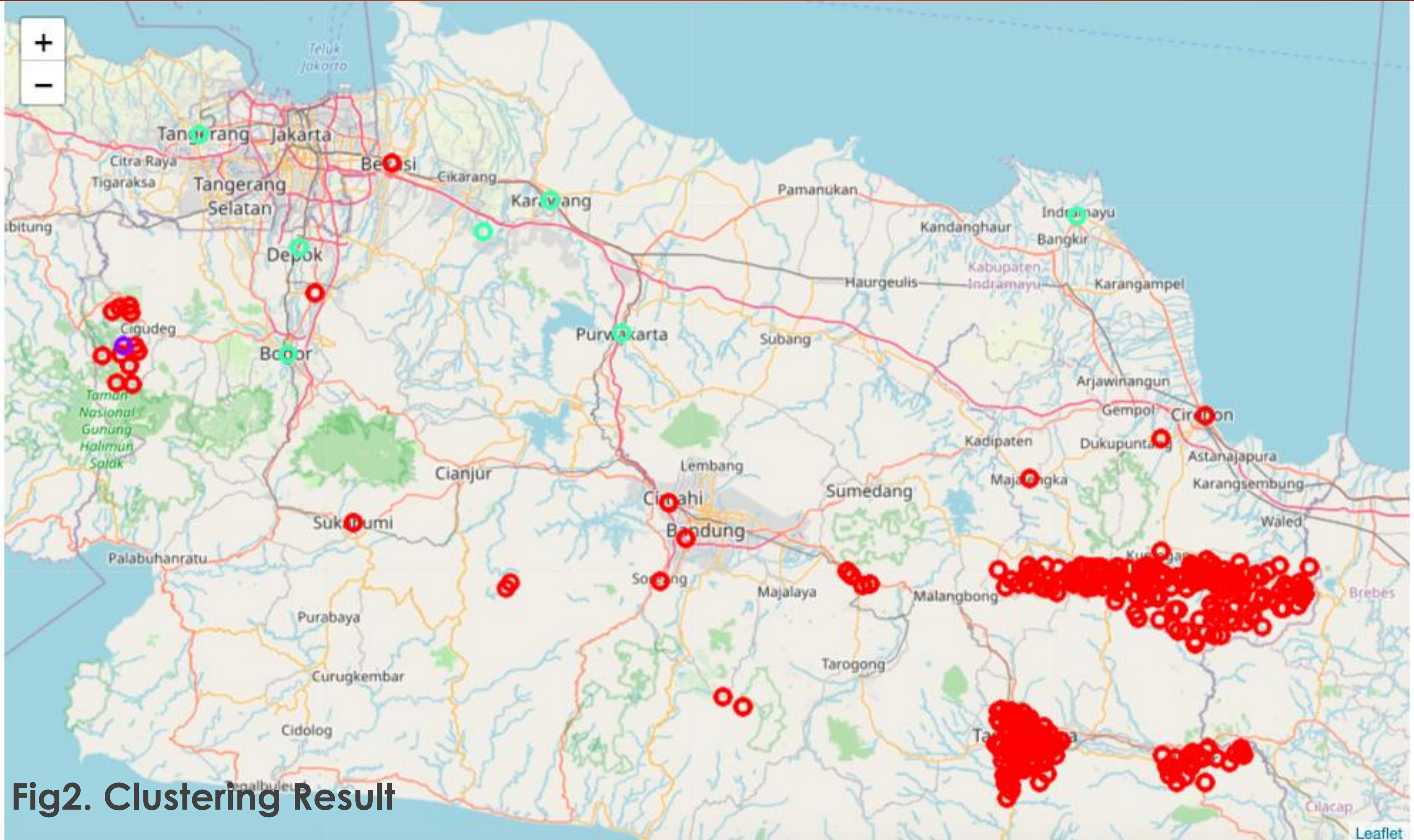
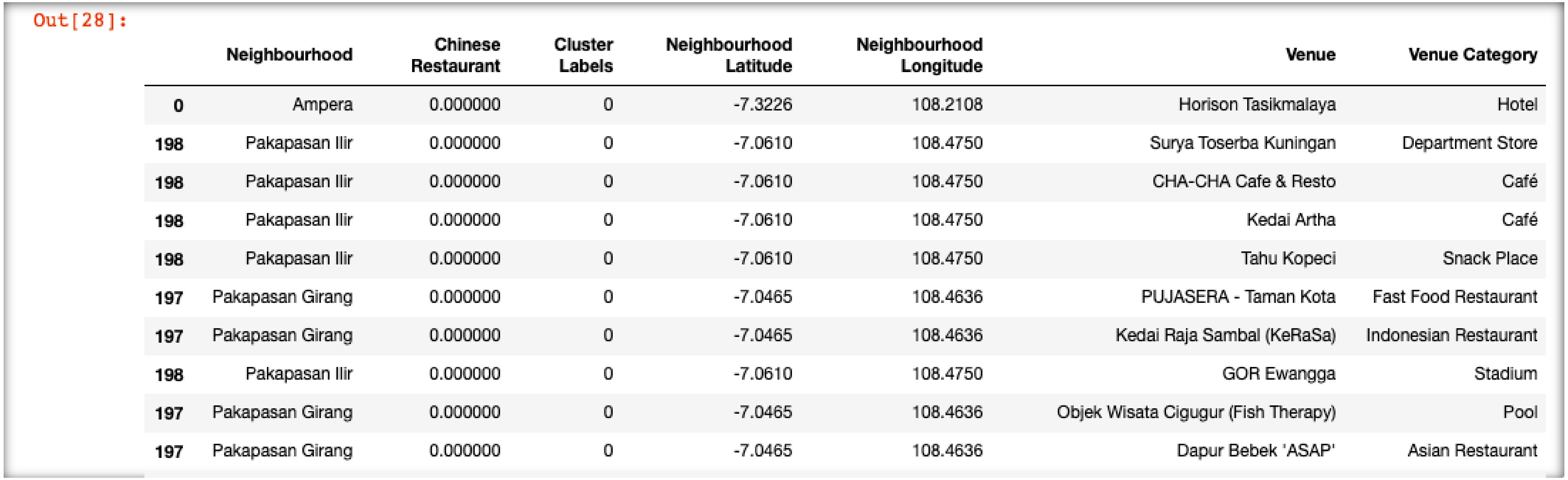


Fig2. Clustering Result

RESULT

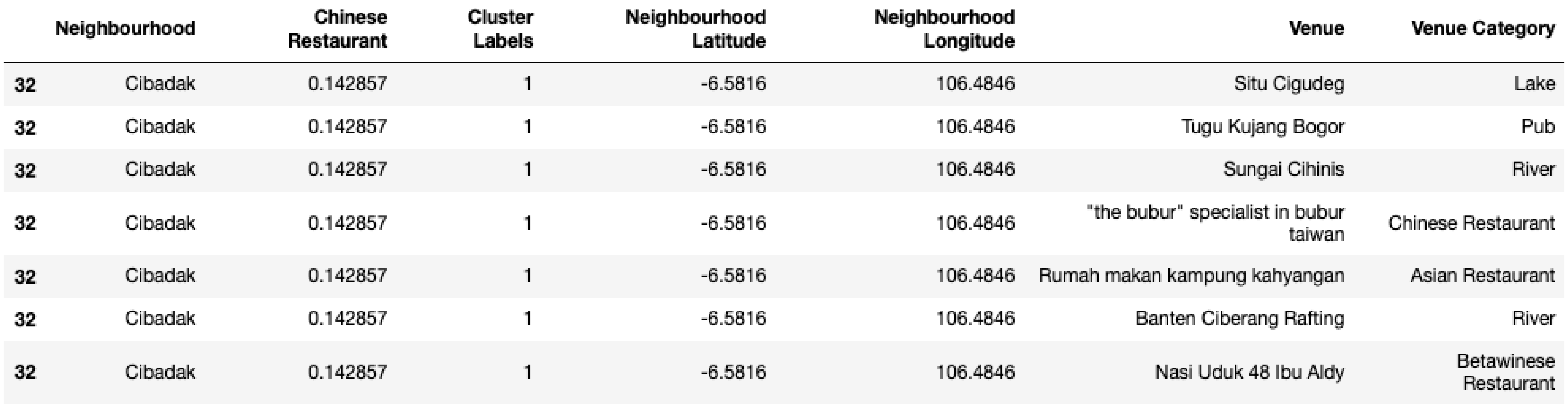


Out[28]:

	Neighbourhood	Chinese Restaurant	Cluster Labels	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Category
0	Ampera	0.000000	0	-7.3226	108.2108	Horison Tasikmalaya	Hotel
198	Pakapasan Ilir	0.000000	0	-7.0610	108.4750	Surya Toserba Kuningan	Department Store
198	Pakapasan Ilir	0.000000	0	-7.0610	108.4750	CHA-CHA Cafe & Resto	Café
198	Pakapasan Ilir	0.000000	0	-7.0610	108.4750	Kedai Artha	Café
198	Pakapasan Ilir	0.000000	0	-7.0610	108.4750	Tahu Kopeci	Snack Place
197	Pakapasan Girang	0.000000	0	-7.0465	108.4636	PUJASERA - Taman Kota	Fast Food Restaurant
197	Pakapasan Girang	0.000000	0	-7.0465	108.4636	Kedai Raja Sambal (KeRaSa)	Indonesian Restaurant
198	Pakapasan Ilir	0.000000	0	-7.0610	108.4750	GOR Ewangga	Stadium
197	Pakapasan Girang	0.000000	0	-7.0465	108.4636	Objek Wisata Cigugur (Fish Therapy)	Pool
197	Pakapasan Girang	0.000000	0	-7.0465	108.4636	Dapur Bebek 'ASAP'	Asian Restaurant

City list on Cluster 0

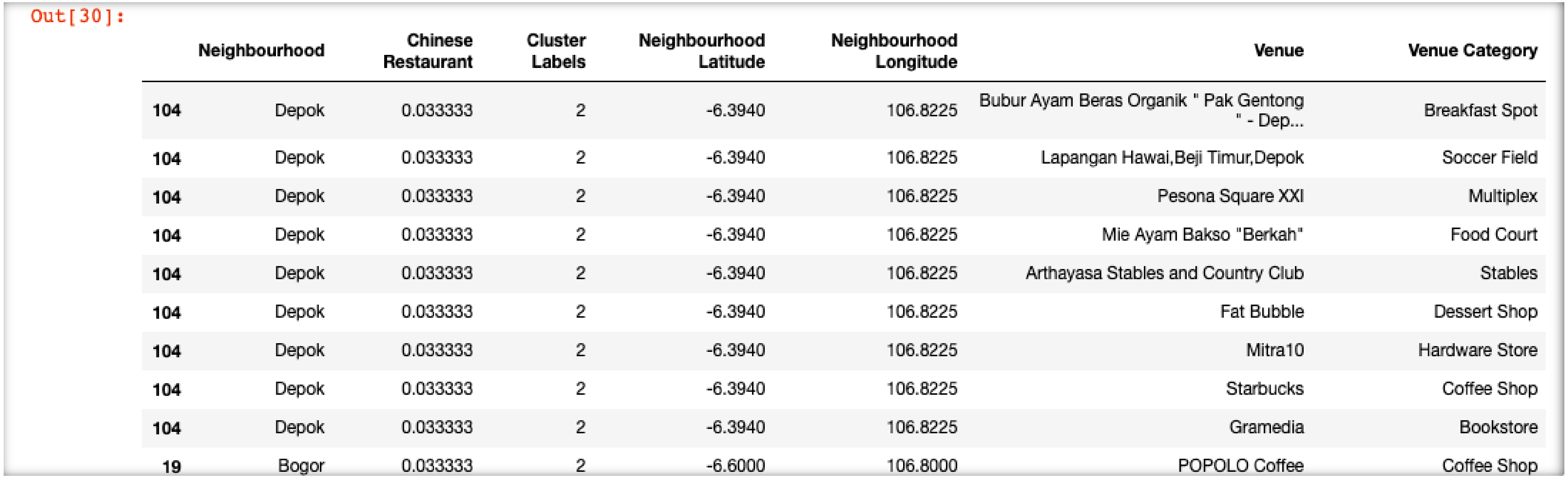
RESULT



	Neighbourhood	Chinese Restaurant	Cluster Labels	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Category
32	Cibadak	0.142857	1	-6.5816	106.4846	Situ Cigudeg	Lake
32	Cibadak	0.142857	1	-6.5816	106.4846	Tugu Kujang Bogor	Pub
32	Cibadak	0.142857	1	-6.5816	106.4846	Sungai Cihinis	River
32	Cibadak	0.142857	1	-6.5816	106.4846	"the bubur" specialist in bubur taiwan	Chinese Restaurant
32	Cibadak	0.142857	1	-6.5816	106.4846	Rumah makan kampung kahyangan	Asian Restaurant
32	Cibadak	0.142857	1	-6.5816	106.4846	Banten Ciberang Rafting	River
32	Cibadak	0.142857	1	-6.5816	106.4846	Nasi Uduk 48 Ibu Aldy	Betawinese Restaurant

City list on Cluster 1

RESULT



Out[30]:

	Neighbourhood	Chinese Restaurant	Cluster Labels	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Category
104	Depok	0.033333	2	-6.3940	106.8225	Bubur Ayam Beras Organik " Pak Gentong " - Dep...	Breakfast Spot
104	Depok	0.033333	2	-6.3940	106.8225	Lapangan Hawai,Beji Timur,Depok	Soccer Field
104	Depok	0.033333	2	-6.3940	106.8225	Pesona Square XXI	Multiplex
104	Depok	0.033333	2	-6.3940	106.8225	Mie Ayam Bakso "Berkah"	Food Court
104	Depok	0.033333	2	-6.3940	106.8225	Arthayasa Stables and Country Club	Stables
104	Depok	0.033333	2	-6.3940	106.8225	Fat Bubble	Dessert Shop
104	Depok	0.033333	2	-6.3940	106.8225	Mitra10	Hardware Store
104	Depok	0.033333	2	-6.3940	106.8225	Starbucks	Coffee Shop
104	Depok	0.033333	2	-6.3940	106.8225	Gramedia	Bookstore
19	Bogor	0.033333	2	-6.6000	106.8000	POPOLO Coffee	Coffee Shop

City list on Cluster 2

Discussion

As the result above, showed that there **three cluster cities** on West Java based on the Chinese Restaurant distribution of each cities. **Cluster 0** is cities who have lowest (or even zero) Chinese Restaurant on it. Mostly the cities are in sub-urban area (except for Bandung). This cities **are recomended** for opening new a Chinese Restaurant (of course other variables are also required to consider, ex : market behavior, city economy outlook, etc. but this is out of this project scope). **Cluster 2** is cities who are in the middle in terms of the number of chinese restaurant. Mostly this cluster is around big cities such as Depok, Bogor, and Karawang. The **recomendation for these cities is very depend on the other variables**. But because mostly on this cluster is big cities, of course we still encouraging people to start their Chinese Restaurant business in here because the market in big cities are also big.

Discussion

Cluster 1 is just one city, it's around Cibadak. This cluster is **not recomended** for opening the Chinese Restaurant because there are a lot of competitor in here. Of course **this recomendation is just based on the number of competitor** side. There are many variables to be considered such as the economy outlook of the city (inflation, unemployent number, etc.) or the market behavior. Nonetheless, this recomendation might be one of the prominent consideration to start a new Chinese Restaurant business in West Java. At least we already know that there are a lot of chinese restaurant around Cibadak, and we don't recomend the city to be the place for opening the Chinese Restaurant.

Conclusion

As the result showed, we recommend the **cities on the clusters 0 to be considered as one of the best location to open a new chinese restaurant.** The consideration must include other variables such as economy outlook in the city (such as GDP, inflation, and unemployment) and market behavior residents on the city. Some of **the cities on cluster 2 are also recommended (around big cities). Cibadak is the only city in cluster 1 which is not recommended for opening the Chinese Restaurant** because there are a lot of competitor in here.