Capstone Project - The Battle of the Neighborhoods (Week 2)

Applied Data Science Capstone by IBM/Coursera

Table of contents

Introduction: Business Problem

In this project we will try to find nearby schools as per finance type. Specifically, this report will be targeted to Realtors working for families (with kids) to find house nearby the school in Hong Kong.

Since there are lots of schools in Hong Kong, we will try to detect **locations that are suitable as per budget**. We would also prefer locations **as close to given location as possible**.

We will use our data science powers to generate a few most promising Districts based on these criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

Data Analysis

Based on definition of our problem, factors that will influence our decision are:

- number of existing schools in the neighborhood.
- type of schools in the neighborhood.

Following data sources will be needed to extract/generate the required information:

- List of Districts will be generated algorithmically and approximate addresses of centers of those areas will be obtained using **Wikipedia link**
- number of schools and their type and location in every district will be obtained using **EDB json file**
- coordinates in Hong Kong will be obtained using **Google Maps API geocoding** of any given location.

Our District Data with longitude and latitude will be look like below:

	District	Income	Latitude	Longitude
0	Hong Kong	20111.764706	22.27832	114.17469
1	Kowloon	30000.000000	22.31667	114.18333
2	Tsuen Wan	32600.000000	22.37137	114.11329
3	Yuen Long Kau Hui	27000.000000	22.45000	114.03333
4	Tung Chung	20111.764706	22.28783	113.94243

Map Visualization for the same.

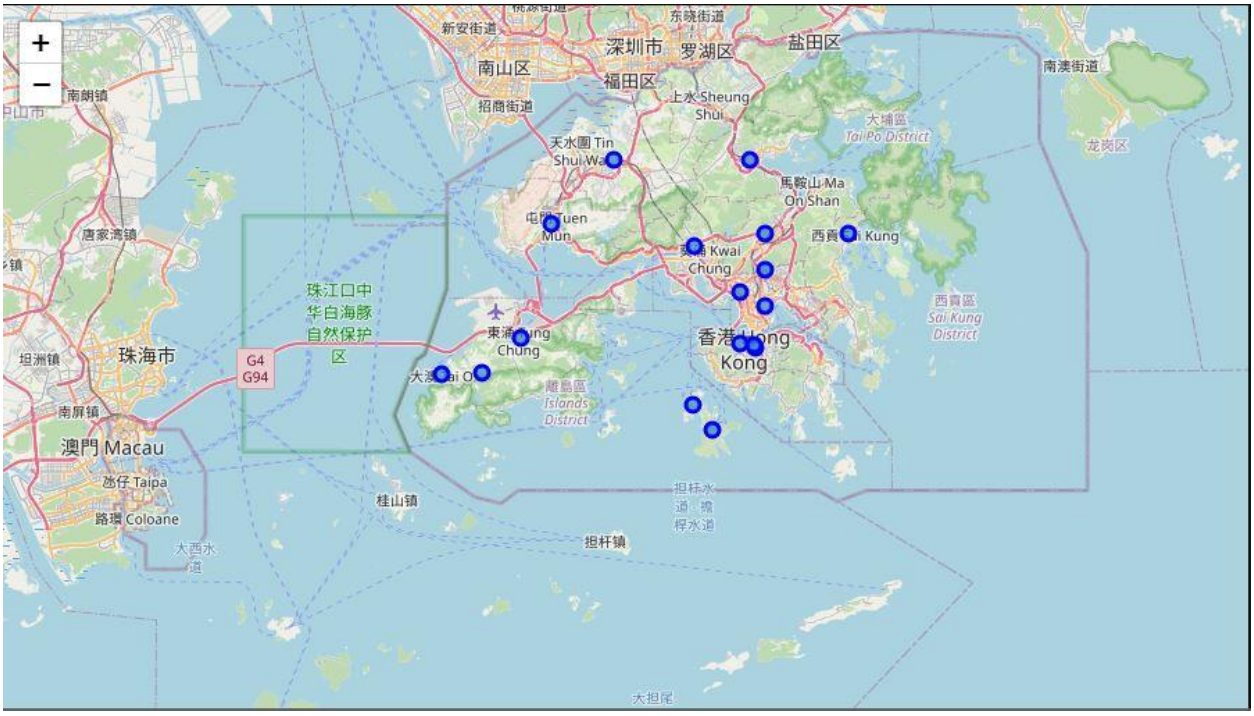


Fig1: Map for districts of Hong Kong.

Methodology

In this project we will direct our efforts on detecting districts of Hong Kong that have schools nearby, particularly those with suitable for all families.

In first step we have collected the required **data: location and type (Finance) of every schools nearby from Tsuen Wan, Hong Kong**.

Second step in our analysis will be calculation and exploration of '**finance Type**' across different areas of Hong Kong - we will use **Bar chart** to identify a few promising areas close to district with different Type of Schools.

In third and final step we will focus on most promising areas and within those create **clusters of locations that meet some basic requirements** established in discussion with stakeholders: we will take into consideration locations with **top 5 schools**. We will present map of all such locations but also create clusters (using **k-means clustering**) of those locations to identify general area / District / location which should be a starting point for final 'street level' exploration and search for optimal school location by families.

Lets Check Number of schools in every area candidate.

	DISTRICT	AIDED	CAPUT	DIRECT SUBSIDY SCHEME	ENGLISH SCHOOLS FOUNDATION	GOVERNMENT	PRIVATE	PRIVATE INDEPENDENT SCH SCHEME
0	YUEN LONG KAU HUI	1	0	0	0	0	0	0
1	YUEN LONG KAU HUI	1	0	0	0	0	0	0
2	YUEN LONG KAU HUI	1	0	0	0	0	0	0
3	YUEN LONG KAU HUI	1	0	0	0	0	0	0
4	YUEN LONG KAU HUI	1	0	0	0	0	0	0

Now Each District with Top 5 Type of Schools:

```
----CENTRAL----
                        school  freq
0                      PRIVATE  0.76
1                        AIDED  0.19
2                   GOVERNMENT  0.02
3        DIRECT SUBSIDY SCHEME  0.01
4  ENGLISH SCHOOLS FOUNDATION  0.01


----KOWLOON----
                        school  freq
0                      PRIVATE  0.74
1                        AIDED  0.20
2        DIRECT SUBSIDY SCHEME  0.02
3                   GOVERNMENT  0.02
4  ENGLISH SCHOOLS FOUNDATION  0.01


----SAI KUNG----
                        school  freq
0                      PRIVATE  0.74
1                        AIDED  0.20
2        DIRECT SUBSIDY SCHEME  0.05
3  ENGLISH SCHOOLS FOUNDATION  0.01
4                   GOVERNMENT  0.01
```

Conversion into Data frame.

	DISTRICT	1st Most Common School	2nd Most Common School	3rd Most Common School	4th Most Common School	5th Most Common School	6th Most Common School	7th Most Common School
0	CENTRAL	PRIVATE	AIDED	GOVERNMENT	DIRECT SUBSIDY SCHEME	ENGLISH SCHOOLS FOUNDATION	CAPUT	PRIVATE INDEPENDENT SCH SCHEME
1	KOWLOON	PRIVATE	AIDED	DIRECT SUBSIDY SCHEME	GOVERNMENT	ENGLISH SCHOOLS FOUNDATION	PRIVATE INDEPENDENT SCH SCHEME	CAPUT
2	SAI KUNG	PRIVATE	AIDED	DIRECT SUBSIDY SCHEME	ENGLISH SCHOOLS FOUNDATION	GOVERNMENT	CAPUT	PRIVATE INDEPENDENT SCH SCHEME
3	SHA TIN	PRIVATE	AIDED	DIRECT SUBSIDY SCHEME	ENGLISH SCHOOLS FOUNDATION	PRIVATE INDEPENDENT SCH SCHEME	GOVERNMENT	CAPUT
4	SHAM SHUI PO	PRIVATE	AIDED	DIRECT SUBSIDY SCHEME	GOVERNMENT	PRIVATE INDEPENDENT SCH SCHEME	CAPUT	ENGLISH SCHOOLS FOUNDATION

Visualize 1 District Data.

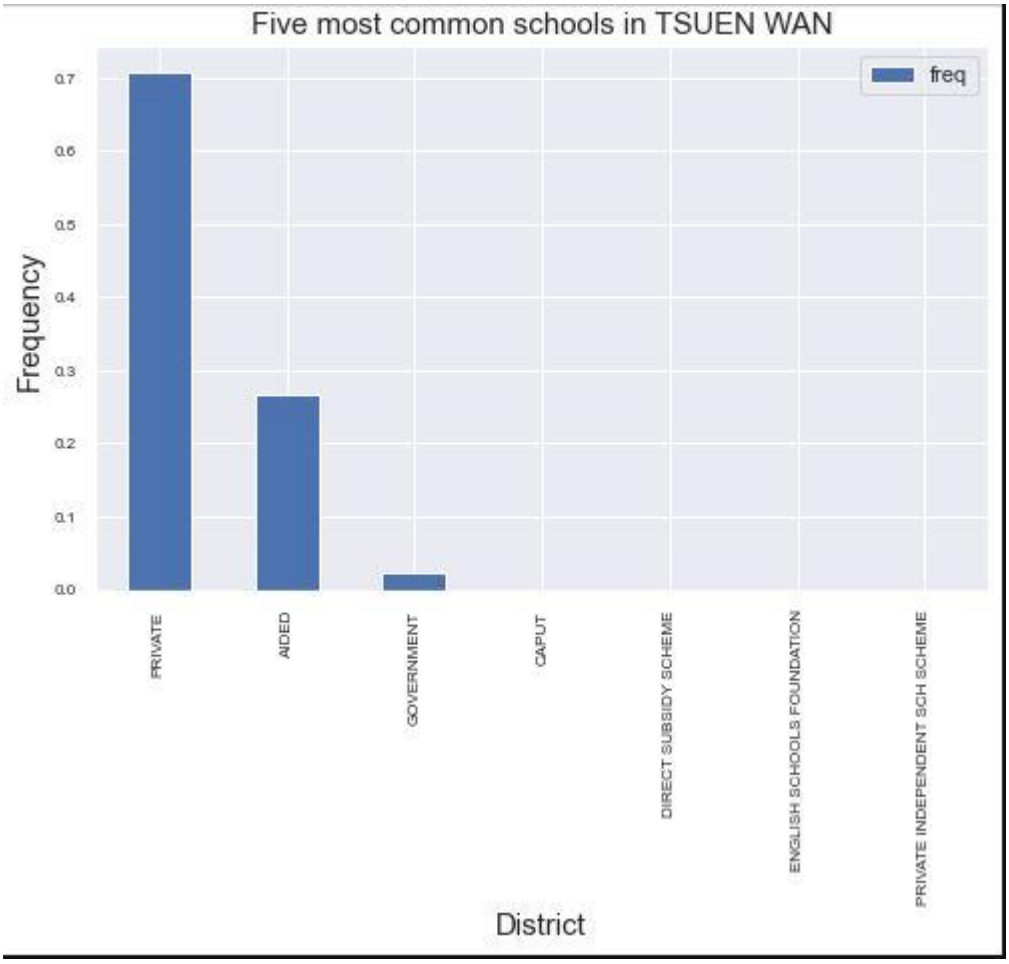
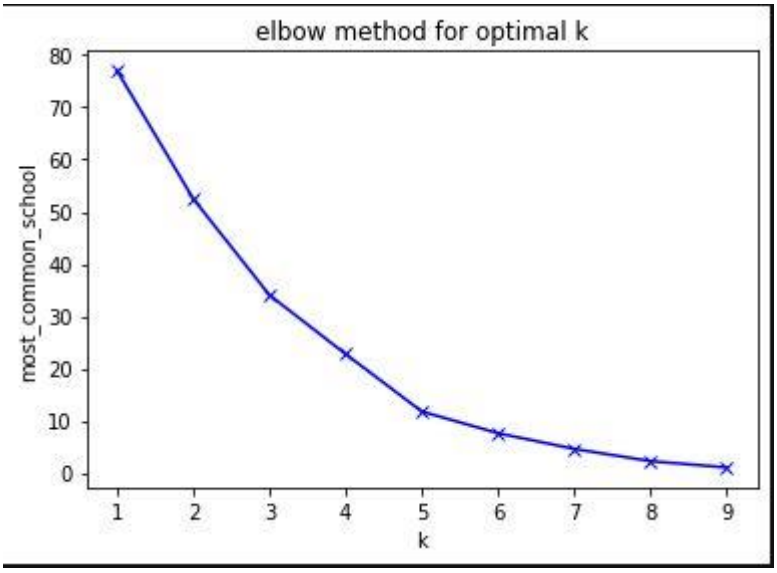


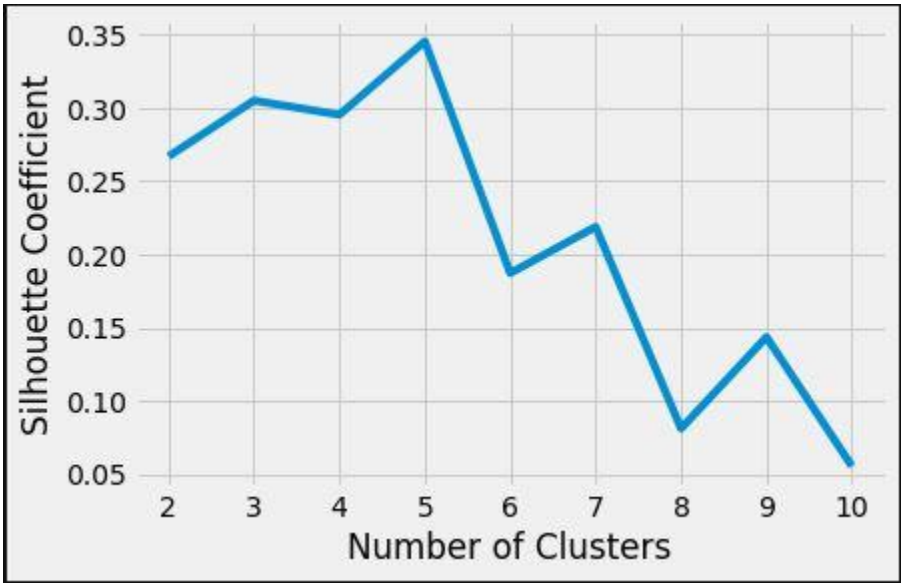
Fig2: Top 7 Schools of Tsuen Wan District to Analyze the Data

Clustering and finding the optimal K means –

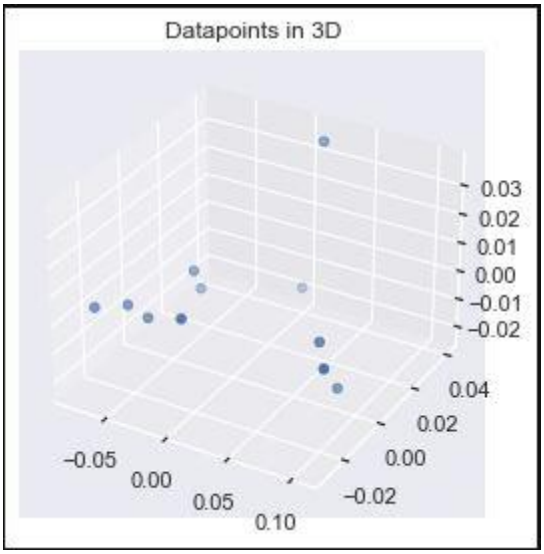
Now we can try to find the option K means by using elbow method.



We can see value of best k value is 5 but to finalize it we can create a plot to confirm our findings.



Now we finalized that k means is 5, Let us visualize our cluster data in 3D diagram.

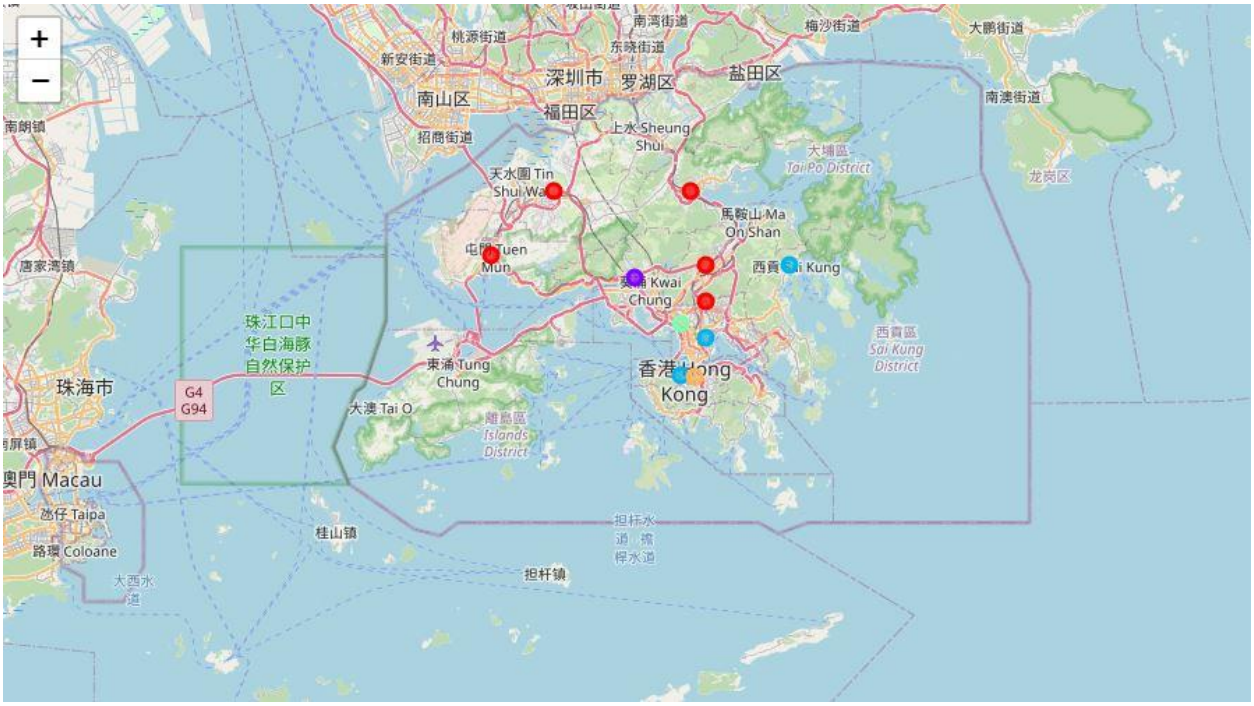


Results and Discussion

Our analysis shows that although there is a great number of schools in Hong Kong, but which are close to nearby districts.

We can see Private, Aided and Government Subsidy schools are most common among all districts and but still we can identify the frequency of popular schools in each district.

We can identify most common schools, and which will help families to identify the schools according to average income.



Conclusion

Purpose of this project was to identify schools nearby to each district of Hong Kong (particularly suitable as per average income) in order to aid stakeholders in narrowing down the search for optimal location for getting the house nearby to school as per the average income.

Clustering of these locations was then performed in order to create major zones of interest (containing greatest number of potential locations) and addresses of those zone centers were created to be used as starting points for final exploration by stakeholders.

We can Define Cluster Names like below:

1) Cluster 1: Lower Medium Average Income Preferred Schools

Cluster 1

```
[53]: hongkong_merge.loc[hongkong_merge['Cluster Labels'] == 0, hongkong_merge.columns[[1] + list(range(5, hongkong_merge.shape[1]))]]
```

[53]:

	INCOME	1st Most Common School	2nd Most Common School	3rd Most Common School	4th Most Common School	5th Most Common School	6th Most Common School	7th Most Common School
3	27000.0	PRIVATE	AIDED	DIRECT SUBSIDY SCHEME	GOVERNMENT	CAPUT	ENGLISH SCHOOLS FOUNDATION	PRIVATE INDEPENDENT SCH SCHEME
5	29700.0	PRIVATE	AIDED	DIRECT SUBSIDY SCHEME	ENGLISH SCHOOLS FOUNDATION	PRIVATE INDEPENDENT SCH SCHEME	GOVERNMENT	CAPUT
6	25000.0	PRIVATE	AIDED	GOVERNMENT	DIRECT SUBSIDY SCHEME	CAPUT	ENGLISH SCHOOLS FOUNDATION	PRIVATE INDEPENDENT SCH SCHEME
7	25800.0	PRIVATE	AIDED	DIRECT SUBSIDY SCHEME	GOVERNMENT	CAPUT	ENGLISH SCHOOLS FOUNDATION	PRIVATE INDEPENDENT SCH SCHEME
16	25500.0	PRIVATE	AIDED	GOVERNMENT	PRIVATE INDEPENDENT SCH SCHEME	DIRECT SUBSIDY SCHEME	CAPUT	ENGLISH SCHOOLS FOUNDATION

2) Cluster 2: Medium Average Income Preferred Schools

Cluster 2

```
[54]: hongkong_merge.loc[hongkong_merge['Cluster Labels'] == 1, hongkong_merge.columns[[1] + list(range(5, hongkong_merge.shape[1]))]]
```

[54]:

	INCOME	1st Most Common School	2nd Most Common School	3rd Most Common School	4th Most Common School	5th Most Common School	6th Most Common School	7th Most Common School
2	32600.0	PRIVATE	AIDED	GOVERNMENT	CAPUT	DIRECT SUBSIDY SCHEME	ENGLISH SCHOOLS FOUNDATION	PRIVATE INDEPENDENT SCH SCHEME

3) Cluster 3: Higher Medium Average Income Preferred Schools

Cluster 3

```
[55]: hongkong_merge.loc[hongkong_merge['Cluster Labels'] == 2, hongkong_merge.columns[[1] + list(range(5, hongkong_merge.shape[1]))]]
```

[55]:

	INCOME	1st Most Common School	2nd Most Common School	3rd Most Common School	4th Most Common School	5th Most Common School	6th Most Common School	7th Most Common School
1	30000.0	PRIVATE	AIDED	DIRECT SUBSIDY SCHEME	GOVERNMENT	ENGLISH SCHOOLS FOUNDATION	PRIVATE INDEPENDENT SCH SCHEME	CAPUT
8	36500.0	PRIVATE	AIDED	DIRECT SUBSIDY SCHEME	ENGLISH SCHOOLS FOUNDATION	GOVERNMENT	CAPUT	PRIVATE INDEPENDENT SCH SCHEME
13	41400.0	PRIVATE	AIDED	GOVERNMENT	DIRECT SUBSIDY SCHEME	ENGLISH SCHOOLS FOUNDATION	CAPUT	PRIVATE INDEPENDENT SCH SCHEME

4) Cluster 4: Lower Average Income Preferred Schools

Cluster 4

```
[56]: hongkong_merge.loc[hongkong_merge['Cluster Labels'] == 3, hongkong_merge.columns[[1] + list(range(5, hongkong_merge.shape[1]))]]
```

[56]:

	INCOME	1st Most Common School	2nd Most Common School	3rd Most Common School	4th Most Common School	5th Most Common School	6th Most Common School	7th Most Common School
15	24300.0	PRIVATE	AIDED	DIRECT SUBSIDY SCHEME	GOVERNMENT	PRIVATE INDEPENDENT SCH SCHEME	CAPUT	ENGLISH SCHOOLS FOUNDATION

5) Cluster 5: Higher Average Income Preferred Schools

Cluster 5

```
[57]: hongkong_merge.loc[hongkong_merge['Cluster Labels'] == 4, hongkong_merge.columns[[1] + list(range(5, hongkong_merge.shape[1]))]]
```

[57]:

	INCOME	1st Most Common School	2nd Most Common School	3rd Most Common School	4th Most Common School	5th Most Common School	6th Most Common School	7th Most Common School
14	44100.0	PRIVATE	AIDED	GOVERNMENT	DIRECT SUBSIDY SCHEME	ENGLISH SCHOOLS FOUNDATION	CAPUT	PRIVATE INDEPENDENT SCH SCHEME

Final decision on optimal school location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone.

Reference

- Hong Kong District Details (https://en.wikipedia.org/wiki/Districts_of_Hong_Kong)
- Hong Kong Districts Geographical Details (https://www.geodatos.net/en/coordinates/hong-kong)
- Hong Kong School geographical Details (https://www.edb.gov.hk/attachment/en/student-parents/sch-info/sch-search/sch-location-info/SCH_LOC_EDB.json)