Coursera Applied Data Science Capstone Project

The Battle for Schools

Ruolin Fan

Introduction

- Hong Kong is a city with extremely high population density and diverse culture. Different districts may possess very different features and environment.
- Education has always been a hot topic and attracts a lot of attentions in HK society. While the educational resources and cultural influences may differ in these different types of schools, the neighborhood around these schools may also have certain patterns that might be influential to the students.
- In an attempt to focus on education in the context of Hong Kong districts, the following questions/aims are raised for this project:

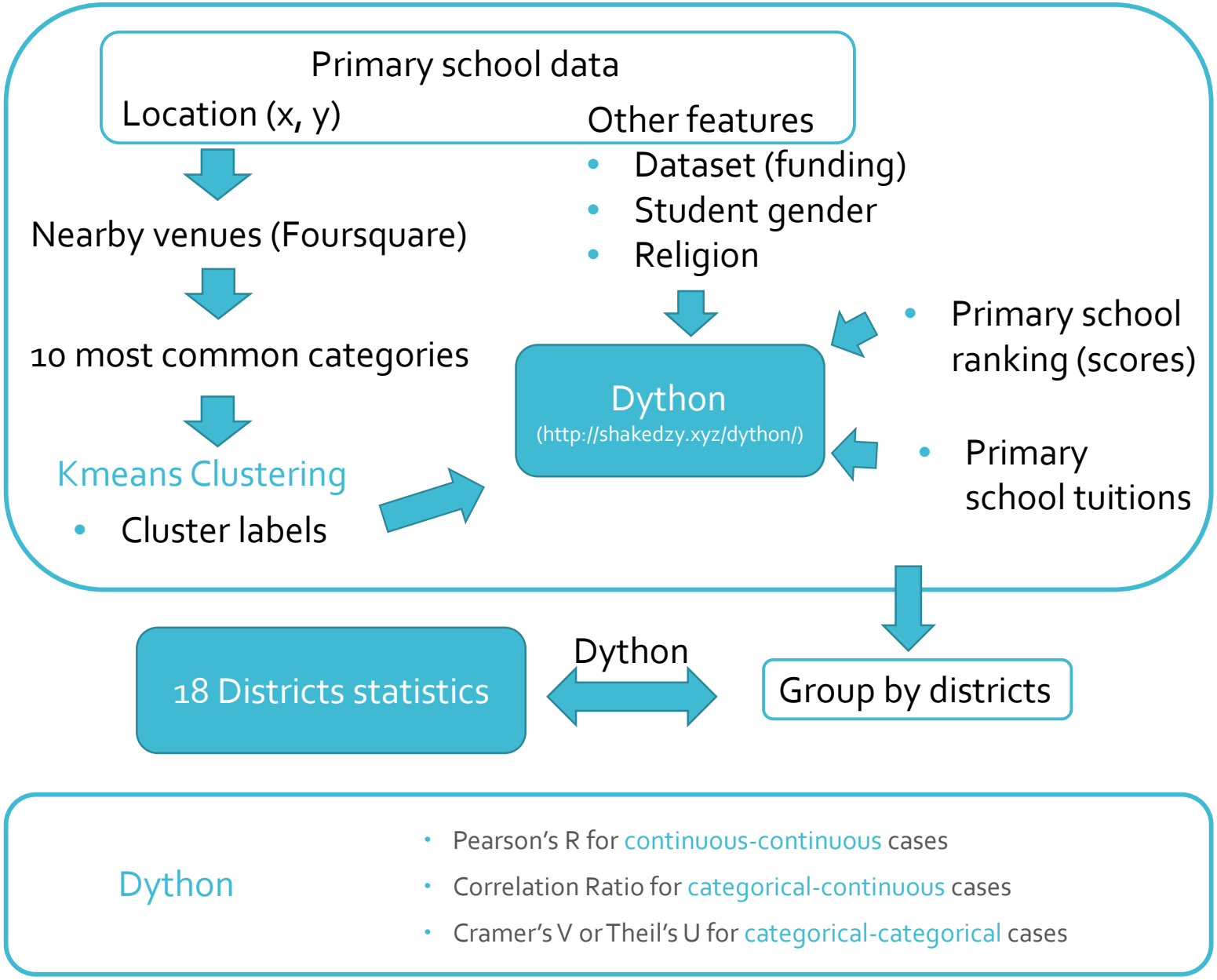
Questions

- Are there any general features for venues nearby primary schools? Can these features help group schools into clusters?
- If the schools can be clustered, does it correlate to its category (government, private, international, etc.?), or the district where it is located?
- Is educational resource correlated with other features of the district, e.g. household income, rent, etc.?
- From the above observations, generate implications for students, parents or educators

Data

- List for primary schools with classified types and location (https://geodata.gov.hk/gs/)
- Ranking of 100 primary schools based on multiple factors.
- Tuition of 54 primary schools.
- Area of 18 districts (https://opendata.esrichina.hk/datasets/eea8ff2f12b145f7b33c4eef4f045513_0/data?orderBy=ENAME)
- Statistics of 18 districts: median monthly domestic household rent, median monthly domestic household income, median mortgage payment and loan repayment, median rent to income ratio, etc. (https://data.gov.hk/en/geospatial-data)

Methodology



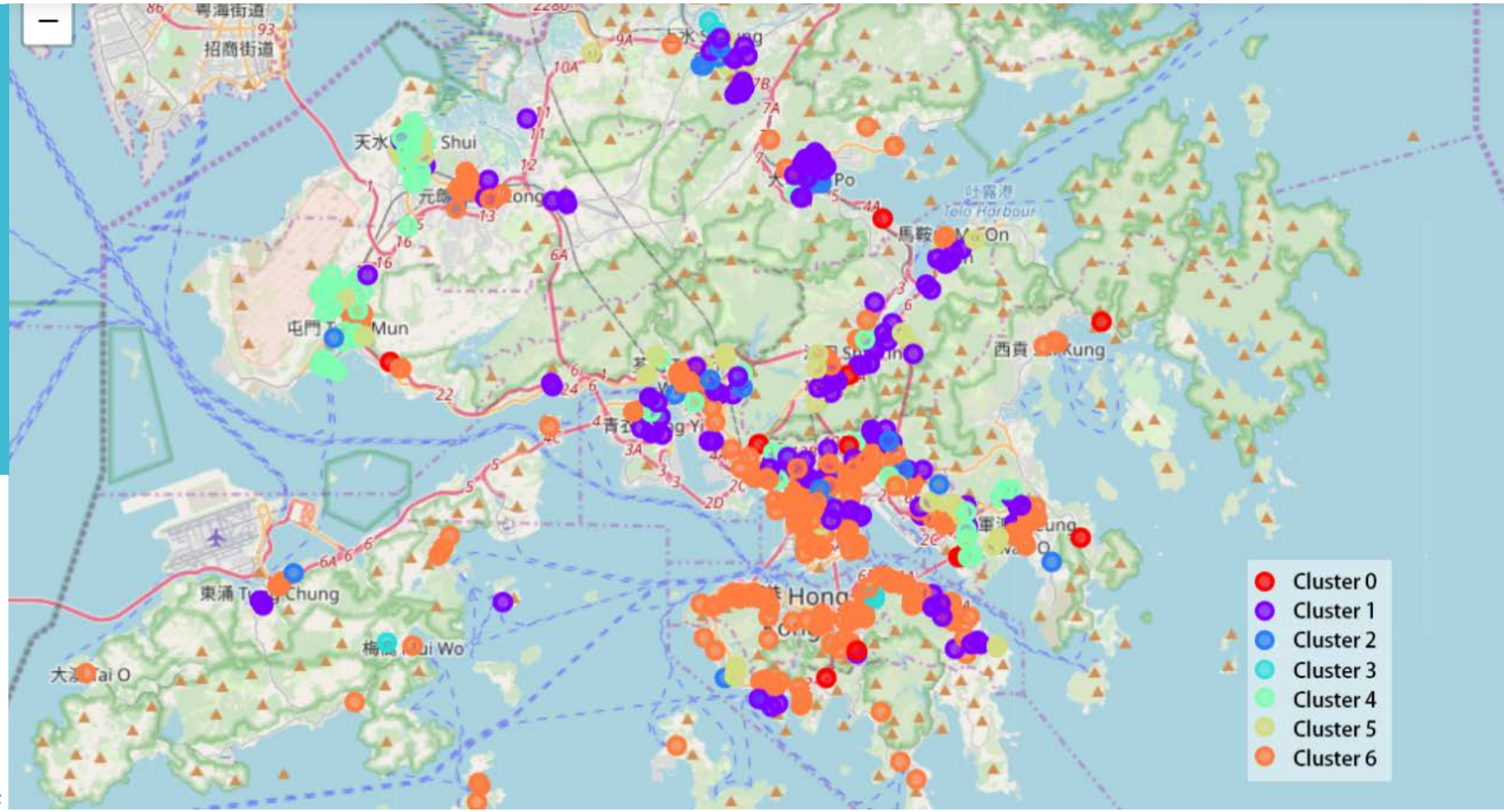
Results

1. Features and distribution of schools

Distribution of schools

-- Classify by clusters --

```
6    242
1    149
4     67
5     56
2     28
0     12
3      5
Name: Cluster Labels, dtype: int64
```



Clustering based on nearby venues

Cluster Labels	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	Chinese Restaurant	Fast Food Restaurant	Hong Kong Restaurant	Shopping Mall	Asian Restaurant	Food Court	French Restaurant	Food Truck	Fried Chicken Joint	Food & Drink Shop
1	Fast Food Restaurant	Chinese Restaurant	Shopping Mall	Hong Kong Restaurant	Park	Cha Chaan Teng	Bus Station	Asian Restaurant	Convenience Store	Food Court
2	Food Court	Food & Drink Shop	Park	Food Truck	Flower Shop	French Restaurant	Chinese Restaurant	Fried Chicken Joint	Fujian Restaurant	Garden
3	Shopping Mall	Light Rail Station	Fast Food Restaurant	Fried Chicken Joint	Food Court	Flower Shop	Food Truck	Food & Drink Shop	Cha Chaan Teng	Flea Market
4	Fast Food Restaurant	Shopping Mall	Hong Kong Restaurant	Cha Chaan Teng	Bus Station	Light Rail Station	Chinese Restaurant	Market	Park	Food Court
5	Coffee Shop	Chinese Restaurant	Café	Food & Drink Shop	Japanese Restaurant	Food Court	Fast Food Restaurant	Food Truck	Flower Shop	Shopping Mall
6	Chinese Restaurant	Coffee Shop	Café	Fast Food Restaurant	Dessert Shop	Cantonese Restaurant	Noodle House	Shopping Mall	Japanese Restaurant	Cha Chaan Teng

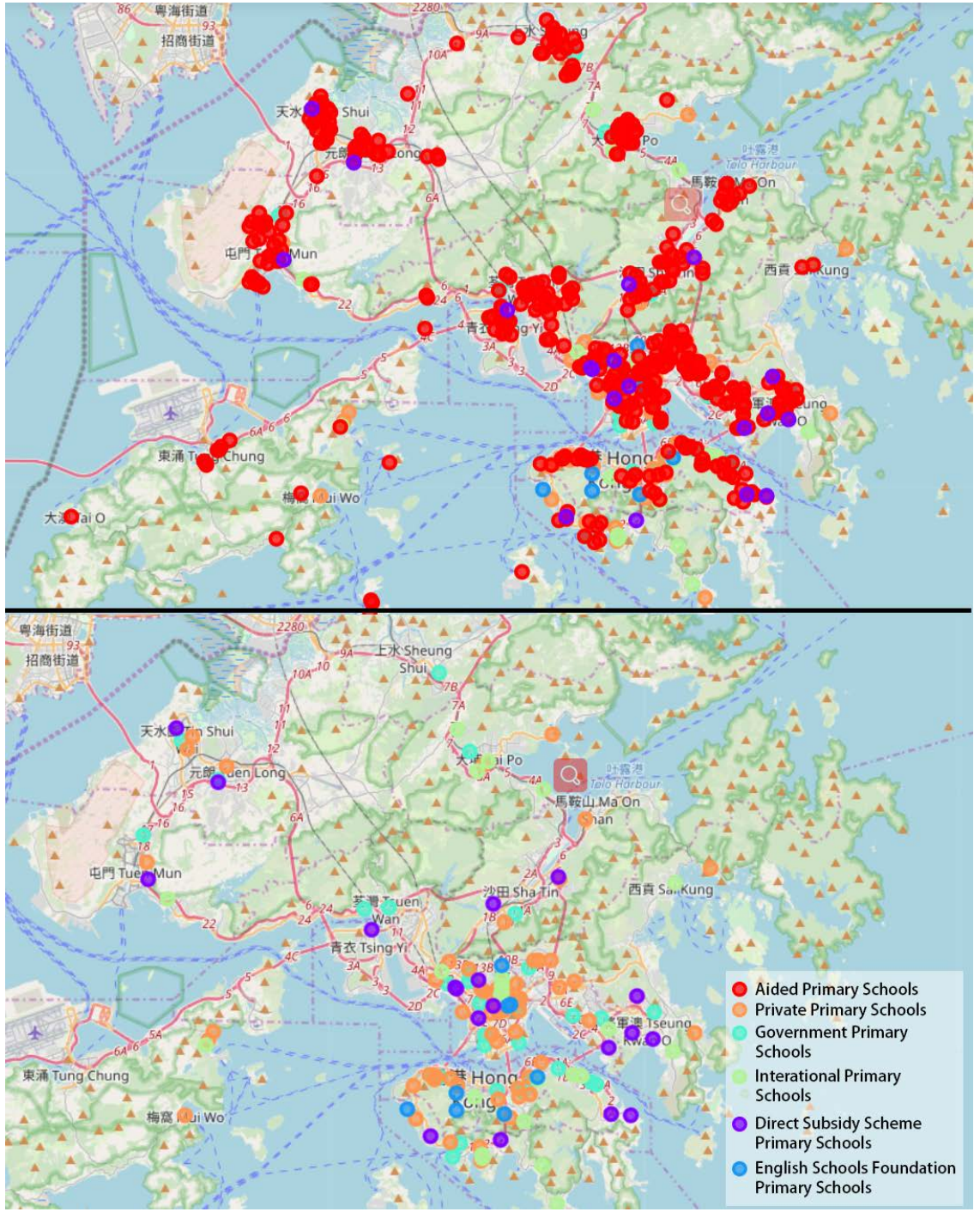
The 10 most common venue categories in each cluster of schools

Distribution of schools

-- Classify by datasets --

Aided Primary Schools	411
Private Primary Schools	64
Government Primary Schools	33
International Schools (Primary)	24
Direct Subsidy Scheme Primary Schools	19
English Schools Foundation (Primary)	8

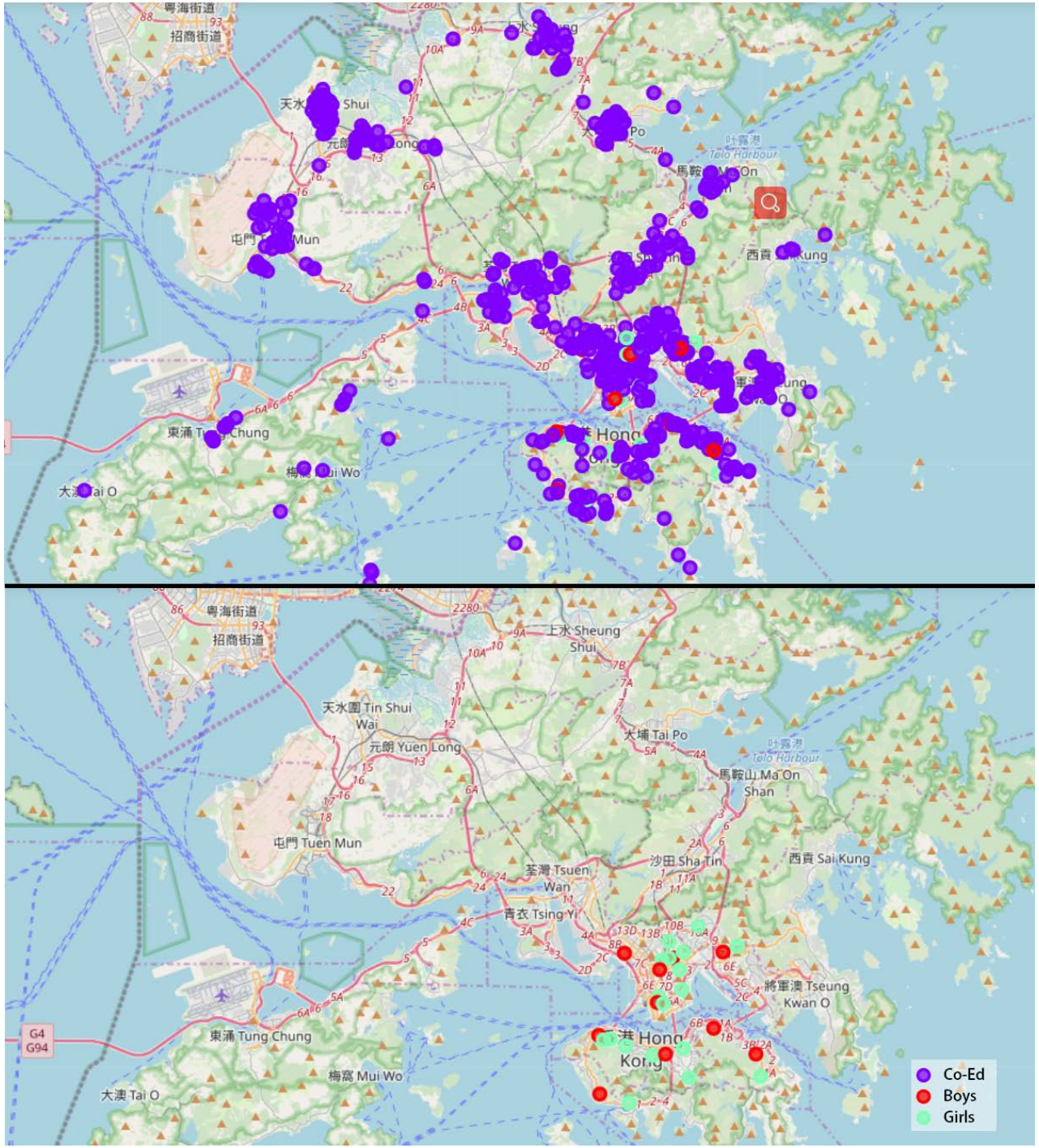
Name: Dataset, dtype: int64



Distribution of schools

-- Classify by student genders --

```
CO-ED     531
GIRLS      18
BOYS       10
Name: Students Gender, dtype: int64
```

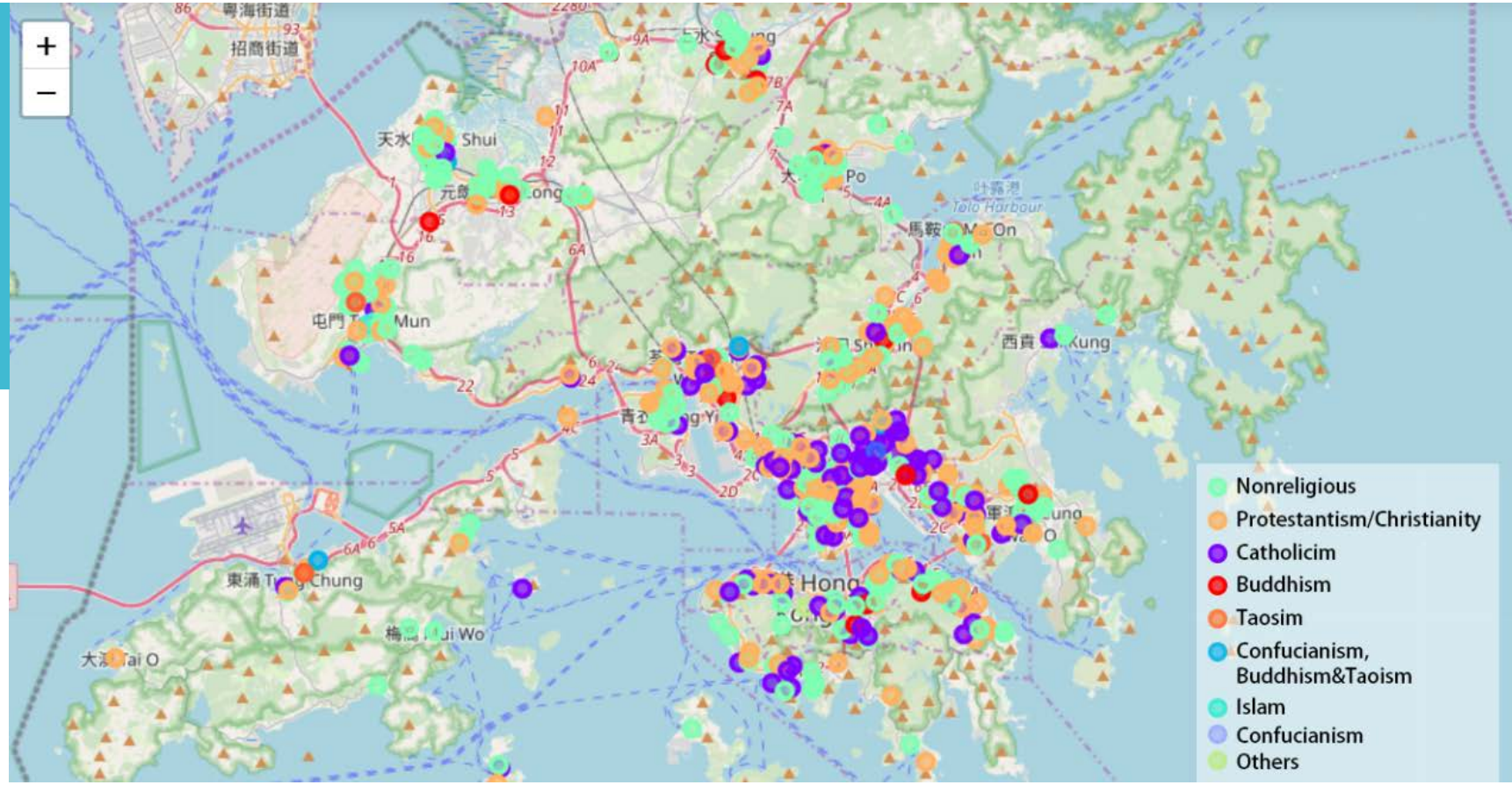


Distribution of schools

-- Classify by religions--

Nonreligious	238
PROTESTANTISM / CHRISTIANITY	183
CATHOLICISM	107
BUDDHISM	15
TAOISM	8
CONFUCIANISM,BUDDHISM & TAOISM	4
ISLAM	2
OTHERS	1
CONFUCIANISM	1

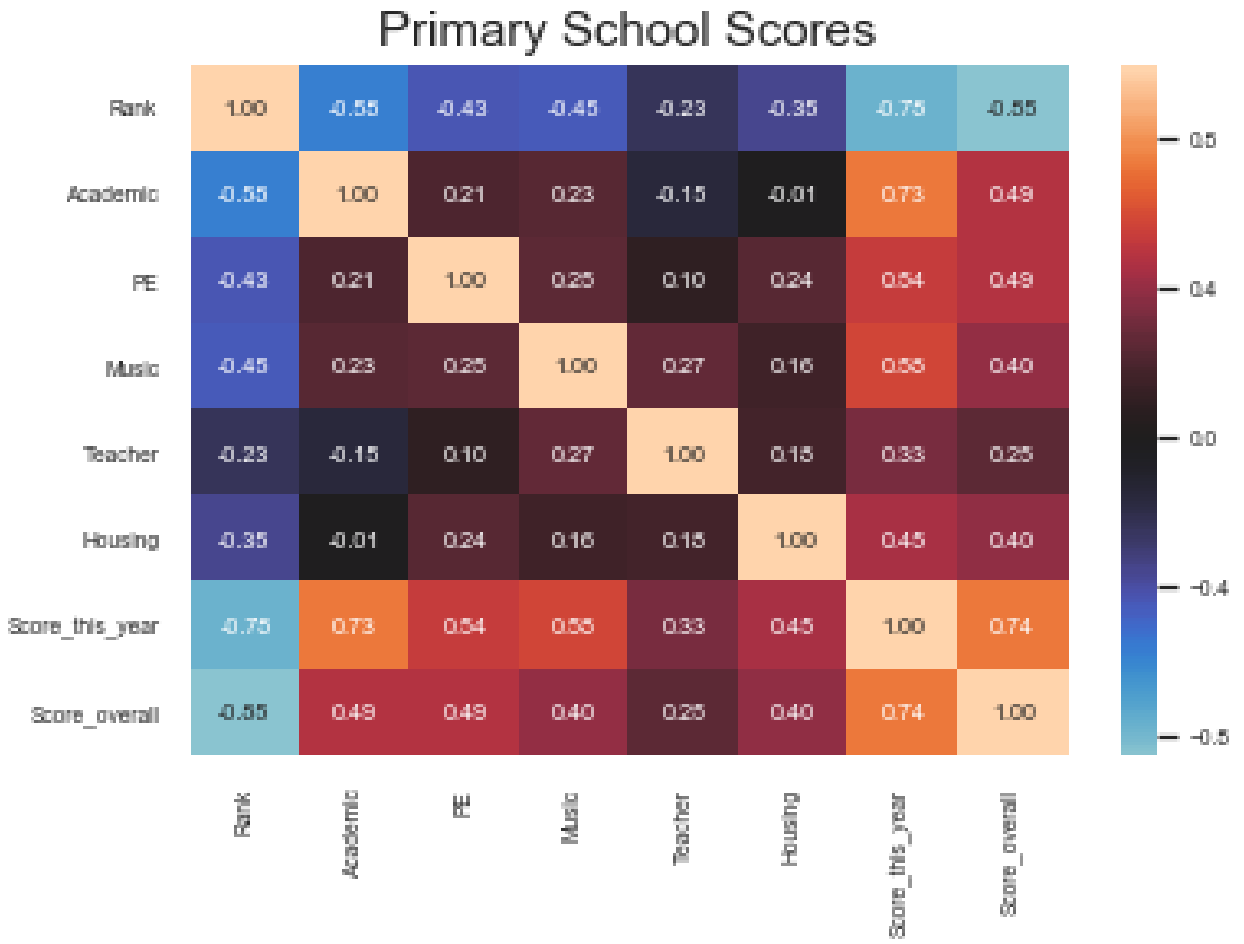
Name: Religion, dtype: int64



Results

2. Scoring of schools and the correlations between evaluating factors

- Ranking has negative correlation with other variables
- The overall score also has a relatively high correlation with other factors, especially the score this year
- Teacher has quite low correlation with all factors but music (excluding Score_this_year or score_overall)
- Housing is most highly correlated with PE
- Teacher and housing scores are even negatively correlated with academic

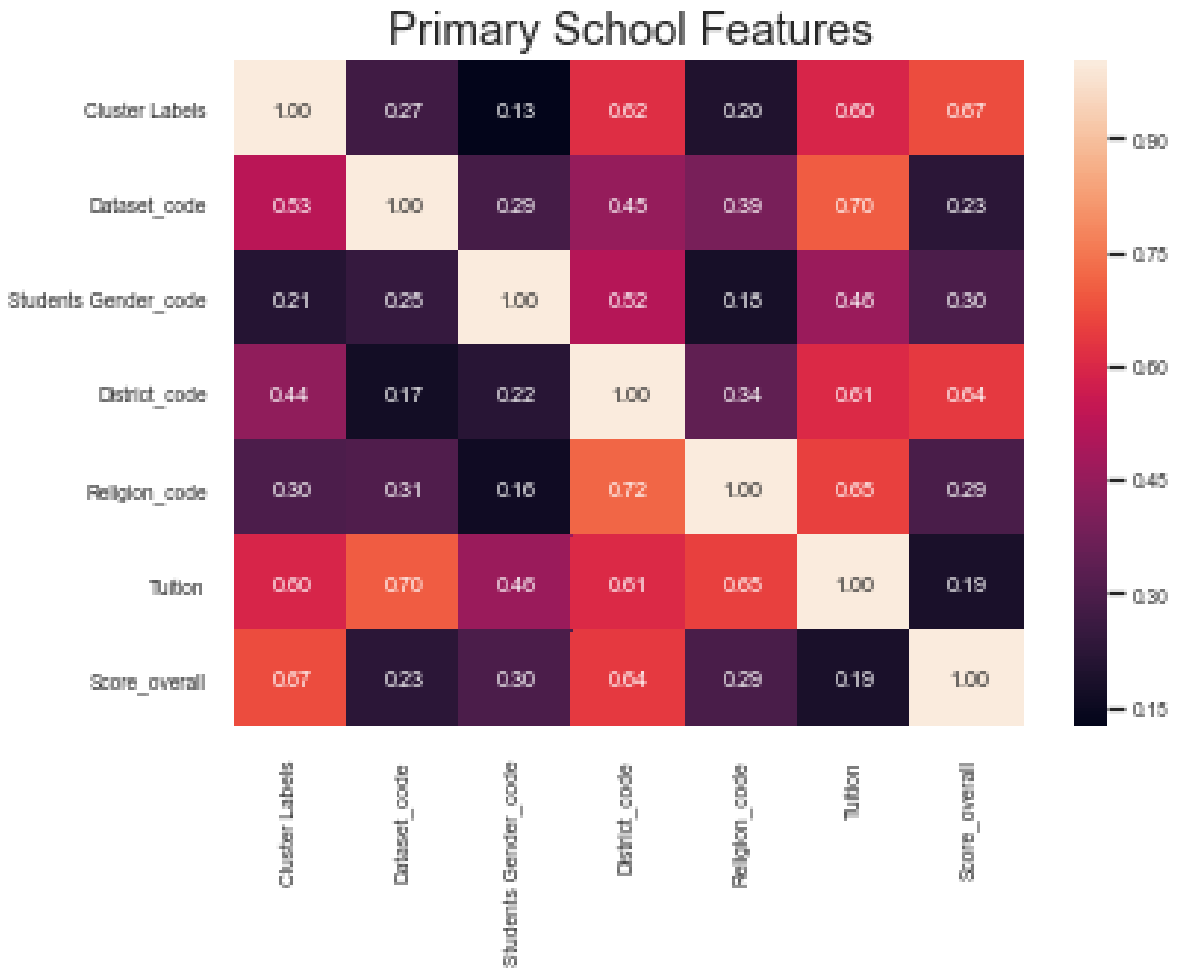


Results

3. Correlations between all features including score and tuition

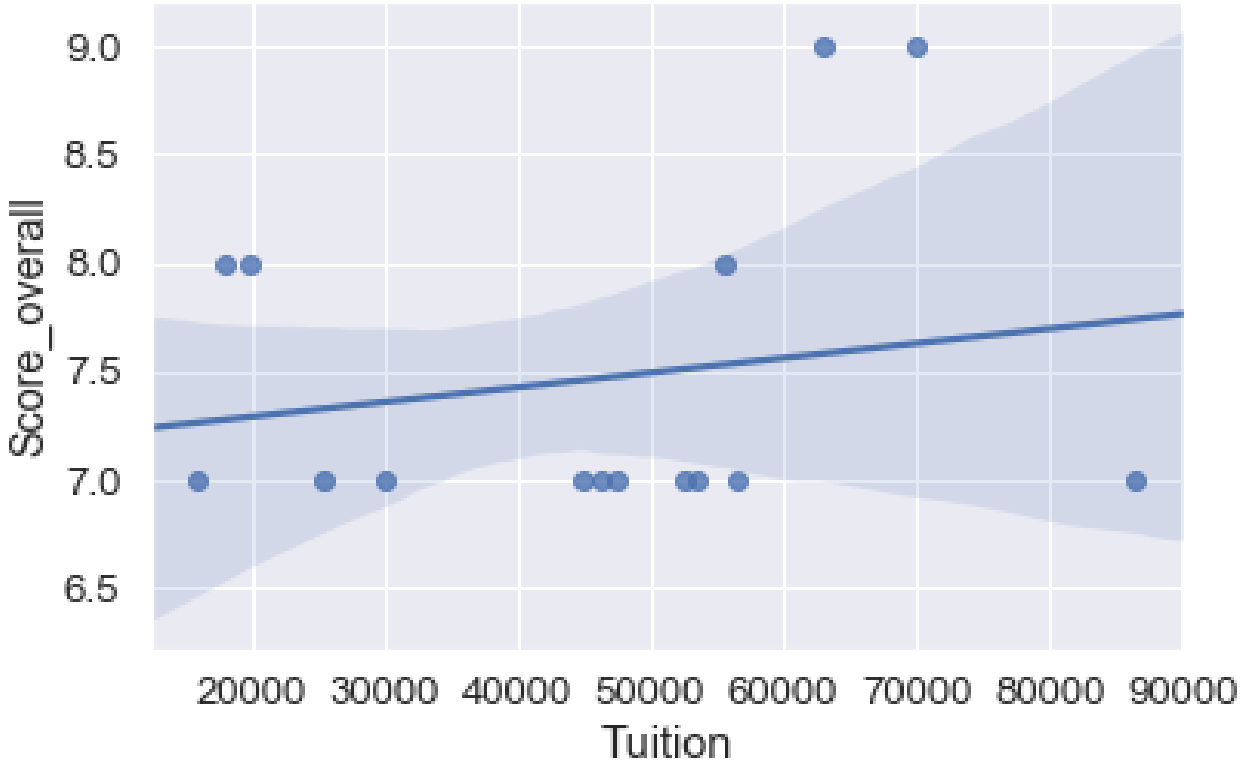
Tuition

- Weak positive correlation between tuition and overall score.
    - Highly scored schools don't necessarily need higher tuition.
    - Small sample size → not representative.
    - Different method for correlation computation → incomparable
- Tuition has highest correlation with dataset
- Other features



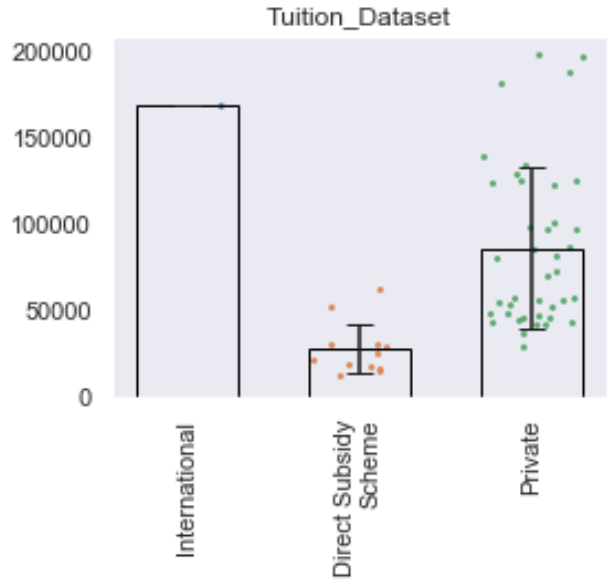
Tuition

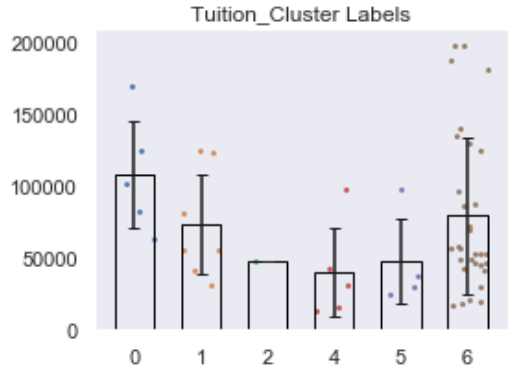
- Weak positive correlation between tuition and overall score.
  - Highly scored schools don't necessarily need higher tuition.
  - Small sample size → not representative.
  - Different method for correlation computation → incomparable
- Tuition has highest correlation with dataset
- Other features

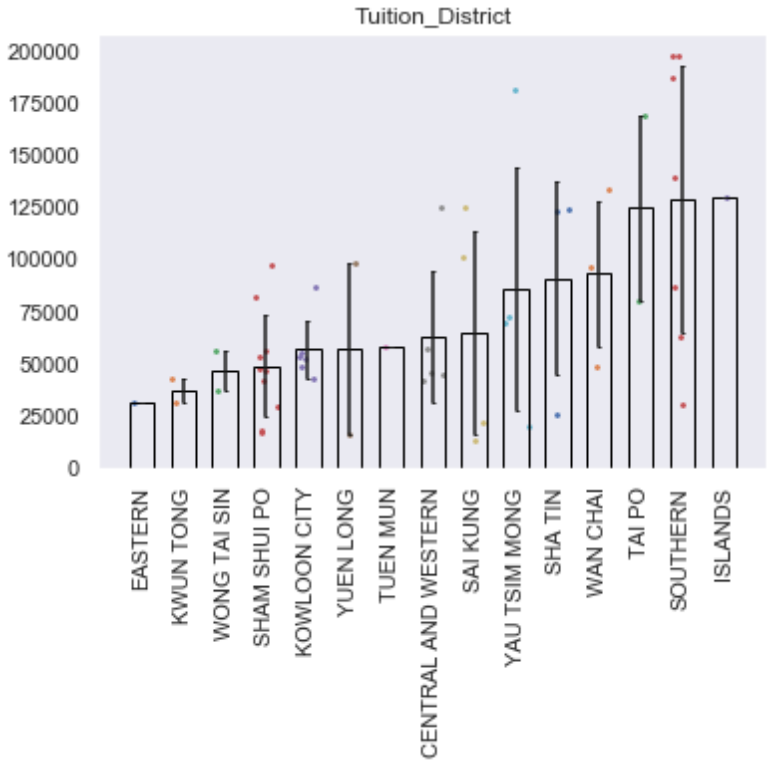


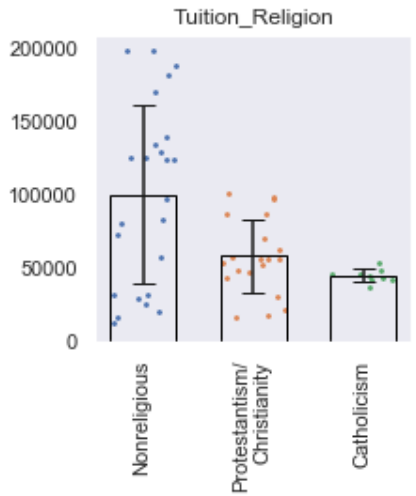
Tuition

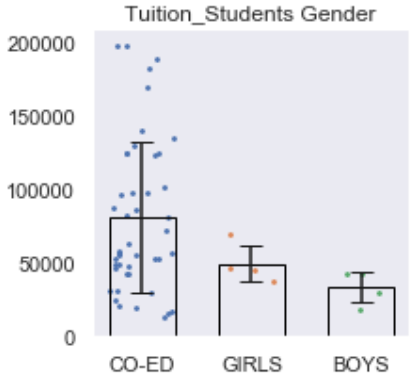
- Weak positive correlation between tuition and overall score.
  - Highly scored schools don't necessarily need higher tuition.
  - Small sample size → not representative.
  - Different method for correlation computation → incomparable
- Tuition has highest correlation with dataset
- Other features





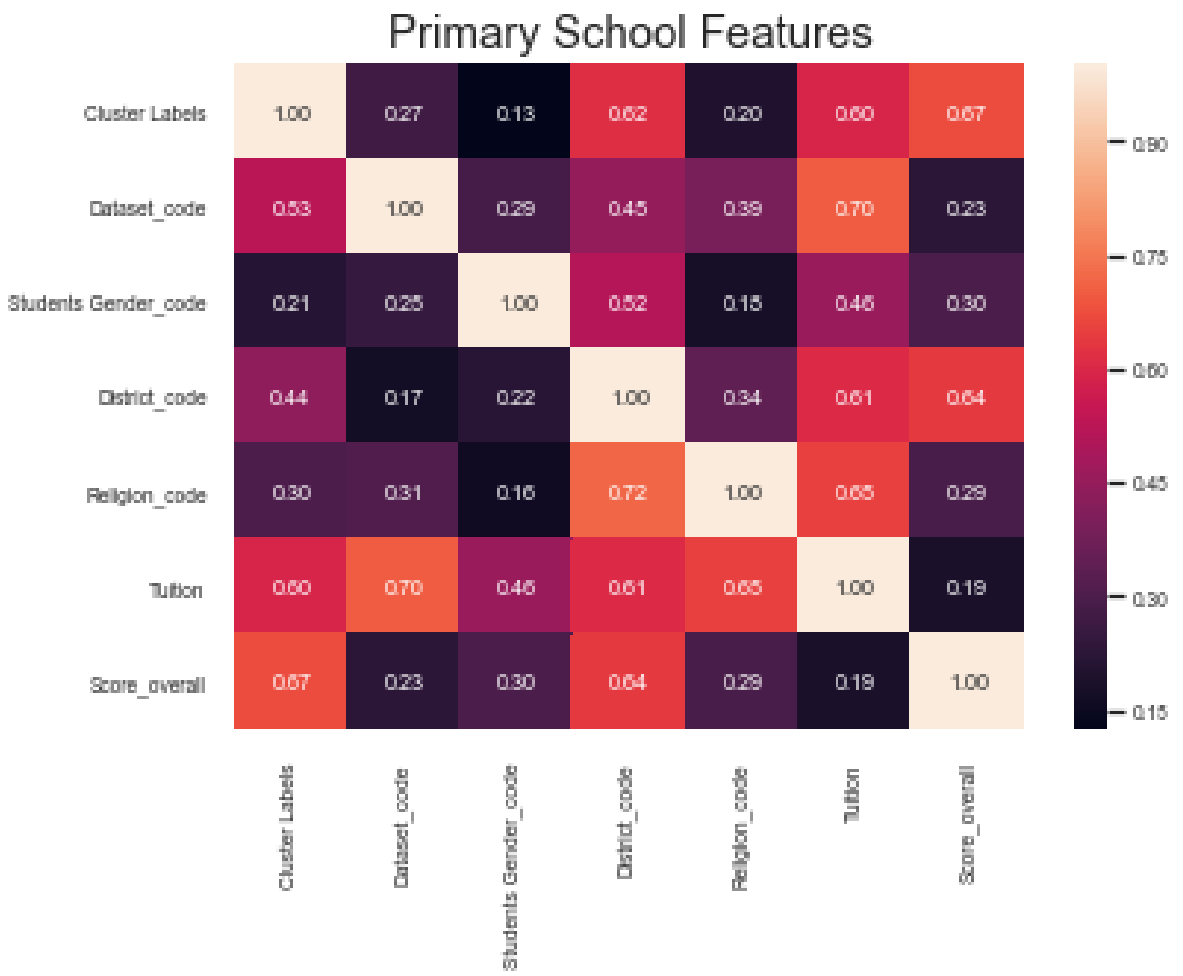






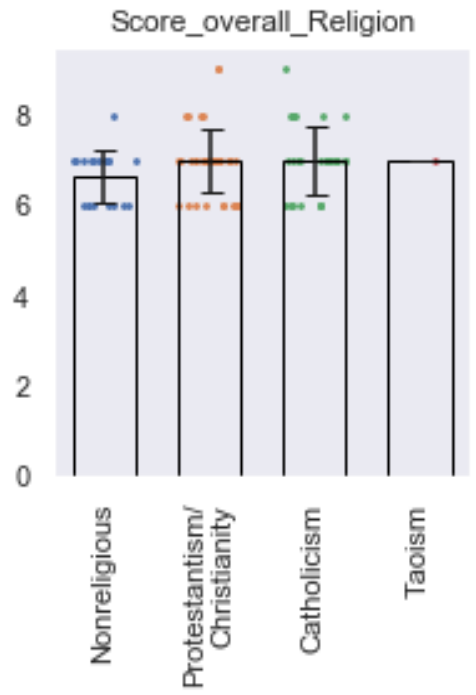
Overall Score

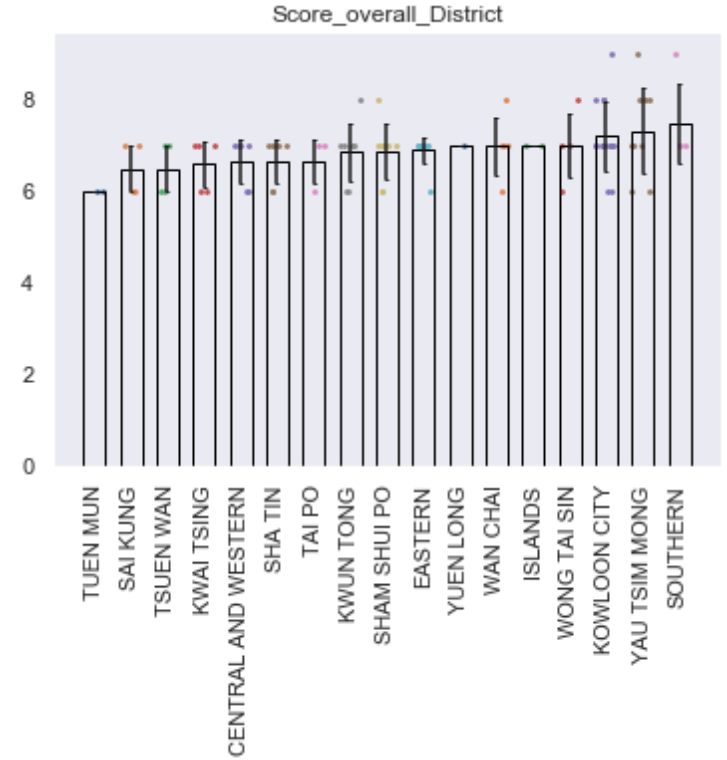
- Overall score is especially highly correlated with cluster label and district
- However, differences between the groups are small.
- A few schools with top scores are located in Kowloon City, Yau Tsim Mong and Southern districts
- Direct subsidy scheme schools and private schools tend to have higher scores
- Single-gender schools tend to have higher score than Co-ed schools.

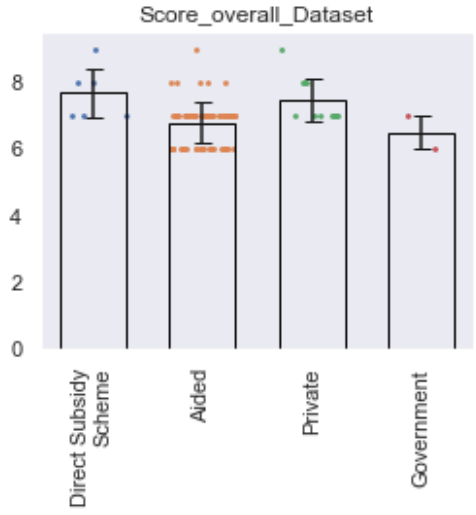


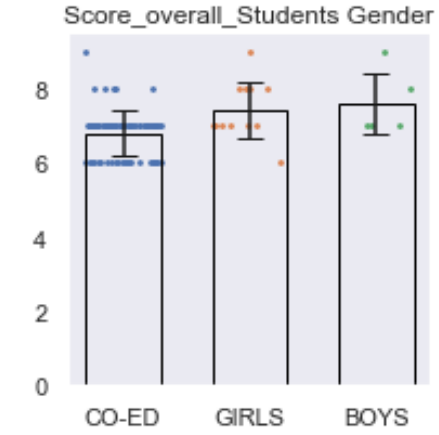
Overall Score

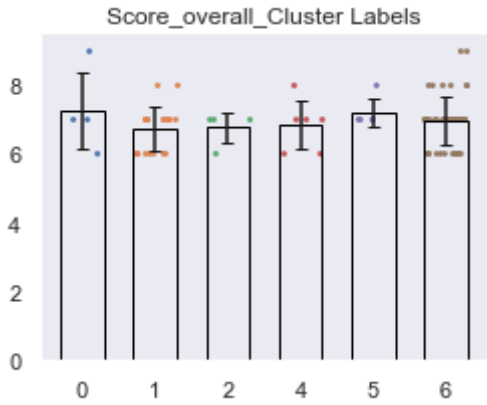
- Overall score is especially highly correlated with cluster label and district
- However, differences between the groups are small.
- A few schools with top scores are located in Kowloon City, Yau Tsim Mong and Southern districts
- Direct subsidy scheme schools and private schools tend to have higher scores
- Single-gender schools tend to have higher score than Co-ed schools.







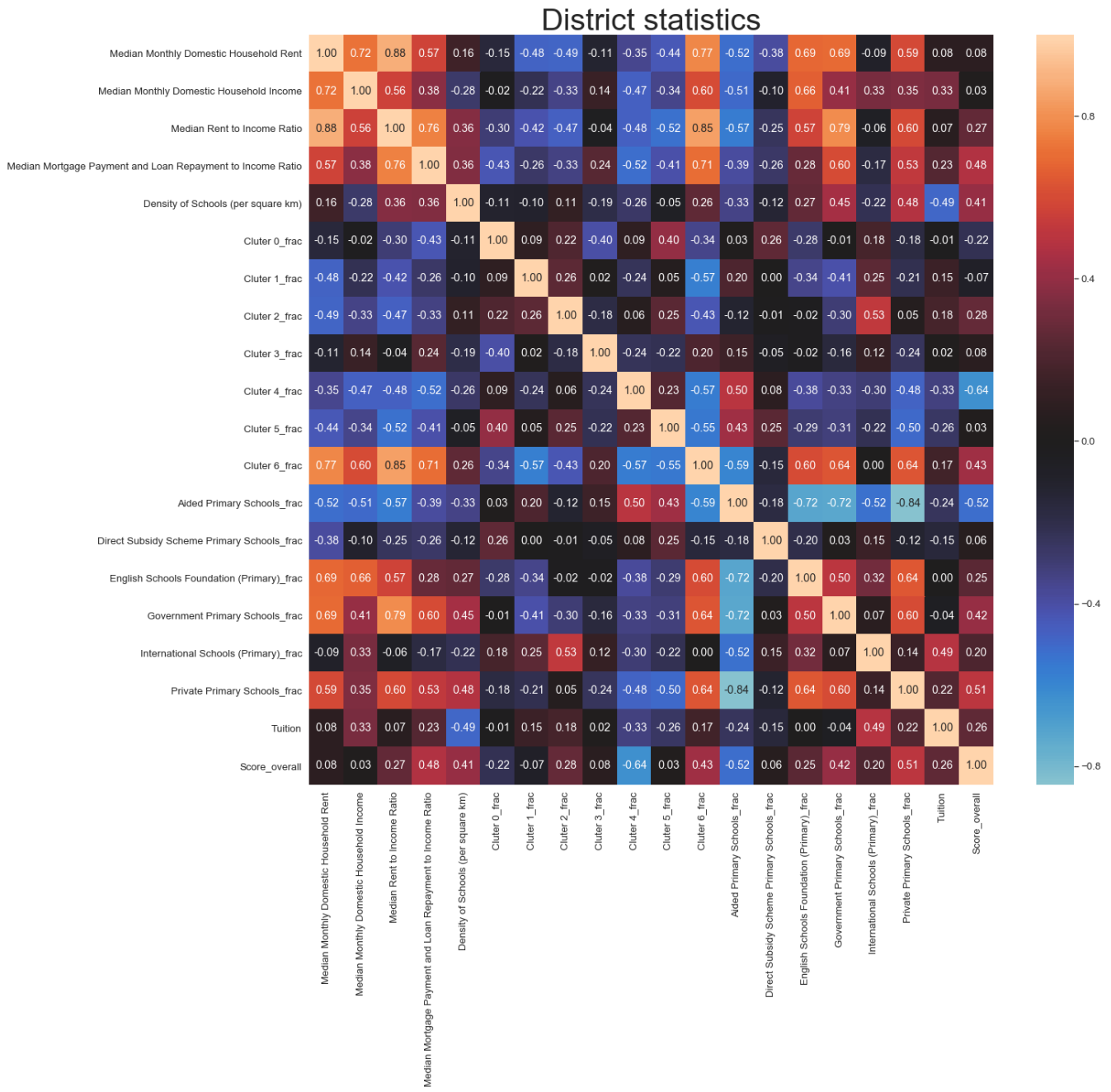




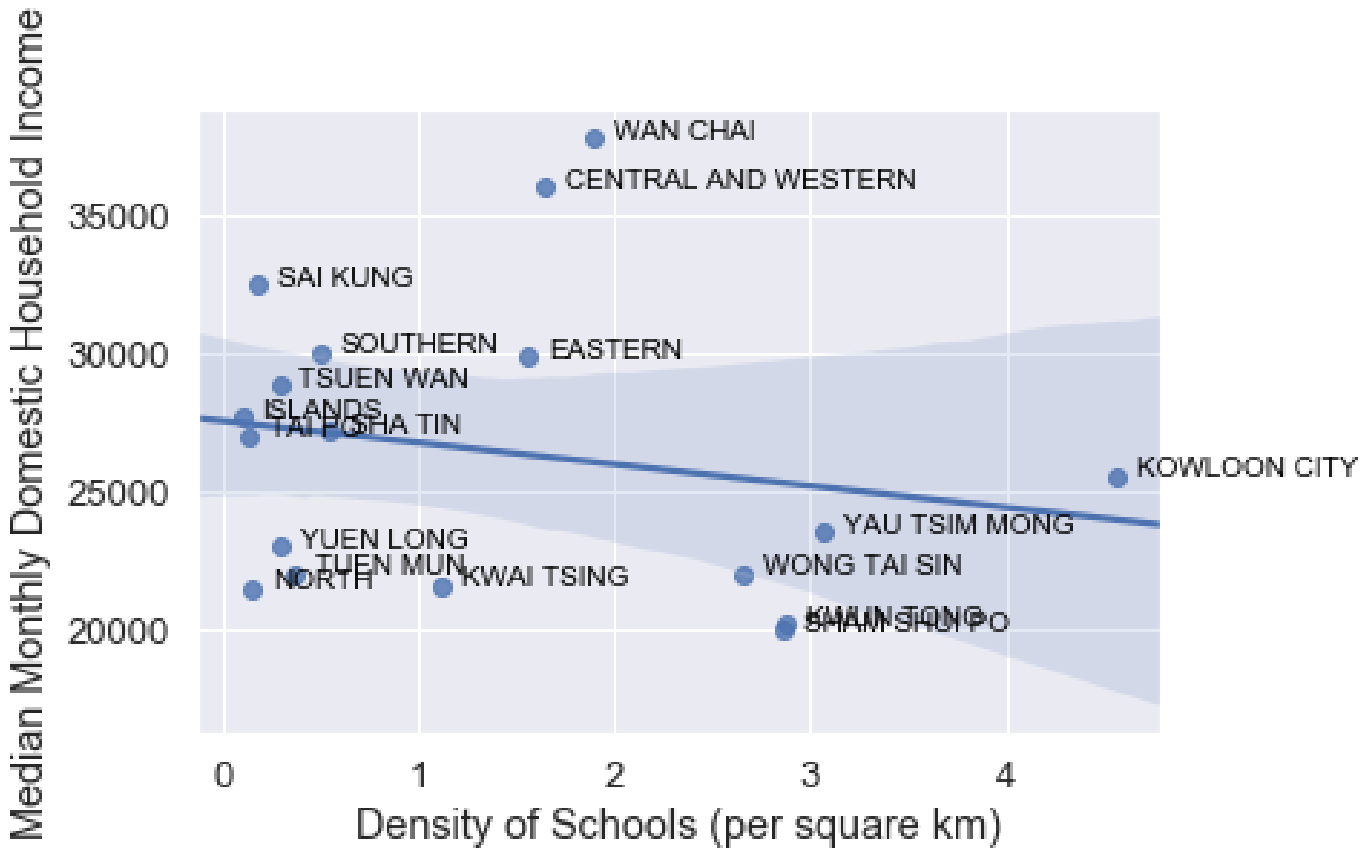
Results

4. Correlations between school distribution in 18 districts and domestic statistics

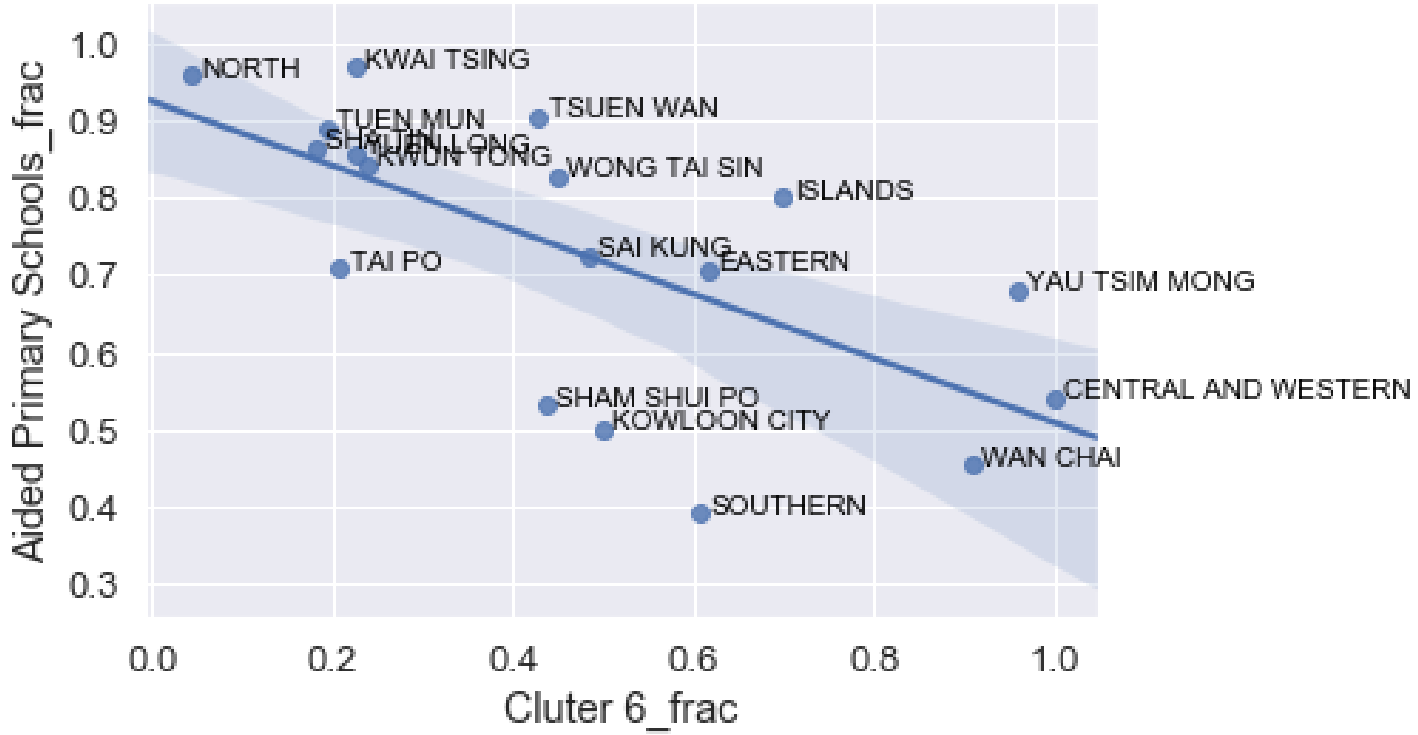
- Four domestic income and rent statistics are positively correlated with each other.
- Only the Median Monthly Domestic Household Income is negatively correlated with the density of schools (with two outliers).
- Cluster 6 and Aided Primary School are the most distinct categories (due to abundance?)
- Overall score is positively correlated with all dataset categories except Aided Primary Schools
- Average tuition has an apparent negative correlation with density of schools (supply-demand relationship).
- A type of districts dominated by cluster 6 schools and correlated with lower proportion of Aided Primary Schools (e.g. Central and Western, Wanchai & Yau Tsim Mong)



- Four domestic income and rent statistics are positively correlated with each other.
- Only the Median Monthly Domestic Household Income is negatively correlated with the density of schools (with two outliers).
- Cluster 6 and Aided Primary School are the most distinct categories (due to abundance?)
- Overall score is positively correlated with all dataset categories except Aided Primary Schools
- Average tuition has an apparent negative correlation with density of schools (supply-demand relationship).
- A type of districts dominated by cluster 6 schools and correlated with lower proportion of Aided Primary Schools (e.g. Central and Western, Wanchai & Yau Tsim Mong)



- Four domestic income and rent statistics are positively correlated with each other.
- Only the Median Monthly Domestic Household Income is negatively correlated with the density of schools (with two outliers).
- Cluster 6 and Aided Primary School are the most distinct categories (due to abundance?)
- Overall score is positively correlated with all dataset categories except Aided Primary Schools
- Average tuition has an apparent negative correlation with density of schools (supply-demand relationship).
- A type of districts dominated by cluster 6 schools and correlated with lower proportion of Aided Primary Schools (e.g. Central and Western, Wanchai & Yau Tsim Mong)



Discussion & Conclusion

Conclusions

- The results revealed a largely similar pool of common venue categories near primary schools, which are mainly restaurants and food shops.
- A certain cluster (cluster 6) is predominant and seems to have a correlation with the most common type of school (aided primary school). The schools in this cluster is especially abundant in Wan Chai, Central and Western, and Yau Tsim Mong districts, which share many similarities in statistics like “Median Rent to Income Ratio”.
- Economic features of a district are correlated with the neighborhood and environment, including the localization of primary schools.

Limitations

- The features used for clustering is essentially not very differentiating and may not be directly related to other features (e.g. tuition and score, etc.).
- Not extracting rating information for these venues for more interesting analysis that might be useful for business purposes
- Single data source for data of tuition and scores that was not evaluated for the credibility due to the limit of resource.

Significance

- Provides an overview about the general features and distributions of primary schools in Hong Kong
- A reference for parents to have an idea about the basic features of each district and the education resources.
- Far from extrapolating any social inferences or decisive conclusions from this simple analysis due to a severe lack of additional information and research.