Victor Alvarado Puyarena
Instructor: Alex Aklson
Applied Data Science Capstone
27 August 2019

Final Report Battle of Schools

Introduction/Business Problem

As a father, the education of your children is one of the most important matters in the family. In Australia the selection of the school or in this case a secondary school is a very important issue for parents, as the system is competitive among the schools located in different suburbs (neighborhoods) in the capital cities of Australia. In many cases, families in Australia take the decision of where to live, based on the school's ranking, the amenities and services available in a determined suburb/neighborhood. For this work, I will develop a

model to classify suburbs in the Melbourne Metro area based on the top secondary schools as per the website betereducation.com.au. The whole purpose of this work is to give a tool to parents, in order to facilitate the decision process that will be driven by data analysis and not just for gut feeling.

"Decisions driven by data are better decisions"

Victor Alvarado P.

Data Description

The data to be use comprises the use of the Foursquare API, where we will use it to gather relevant information about venues surrounding the top secondary schools in Melbourne. For that end we will be using the data from the website www.betereducation.com.au, where I will scrap a table with the top secondary schools in Victoria. On that table there is information related to the address of the school, however, there some schools have no information of the suburb and also the postcode could be share among several suburbs as well.

Dataframe with the web scrapped information from:
https://bettereducation.com.au/school/secondary/vic/vic_top_secondary_schools.aspx

	School	Postcode	State Overall Score	Total Enrolments	Sector
0	Mac.Robertson Girls' High School	3004	100	954	Government
1	Melbourne High School	3141	100	1357	Government
2	Nossal High School,Berwick,VIC,3806	3806	100	832	Government
3	Suzanne Cory High School,Werribee,VIC,3030	3030	100	871	Government
4	Presbyterian Ladies' College,Burwood,VIC,3125	3125	100	1407	Non-government

Another source of data came from https://en.wikipedia.org/wiki/List_of_Melbourne_suburbs where it was imported a dataset with the Postcodes of the Victoria state in Australia from.

After we scrap that information into a pandas' dataframe, we found that Google Sheets was able to geocode a massive amount of address with a very stable performance. Below, a dataframe with suburbs postcodes and its relative coordinates in Latitude and Longitude.

	Suburb	Postcode	State	Address	Lat	Lng	Local government area
0	Melbourne CBD	3000	VIC	Melbourne CBD,3000,VIC	-37.815207	144.963937	City of Melbourne
1	East Melbourne	3002	VIC	East Melbourne,3002,VIC	-37.816144	144.980459	City of Melbourne
2	West Melbourne	3003	VIC	West Melbourne,3003,VIC	-37.811450	144.925397	City of Melbourne
3	Melbourne CBD(St Kilda Roadarea)	3004	VIC	Melbourne CBD(St Kilda Roadarea),3004,VIC	-37.838947	144.975959	City of Melbourne(east side to High Street, Pr...
4	South Wharf	3006	VIC	South Wharf,3006,VIC	-37.825000	144.952000	City of Melbourne;City of Port Phillip

Using the information from the above table, the school dataset was filled with the missing information about the address of each school, that is based on School's name, suburb name, state code (in this case VIC for Victoria) and postcode.

	School	Postcode	State Overall Score	Total Enrolments	Sector
0	Mac.Robertson Girls' High School,Melbourne CBD...	3004	100	954	Government
1	Melbourne High School,South Yarra,VIC,3141	3141	100	1357	Government
2	Nossal High School,Berwick,VIC,3806	3806	100	832	Government
3	Suzanne Cory High School,Werribee,VIC,3030	3030	100	871	Government
4	Presbyterian Ladies' College,Burwood,VIC,3125	3125	100	1407	Non-government
5	Fintona Girls' School,Balwyn,VIC,3103	3103	100	450	Non-government
6	Haileybury College,Keysborough	3173	99	3927	Non-government
7	Huntingtower School,Mount Waverley,VIC,3149	3149	99	697	Non-government

With the 'School' Column in a Google Sheet, we proceeded to geocode all the schools to obtain the geographical coordinates.

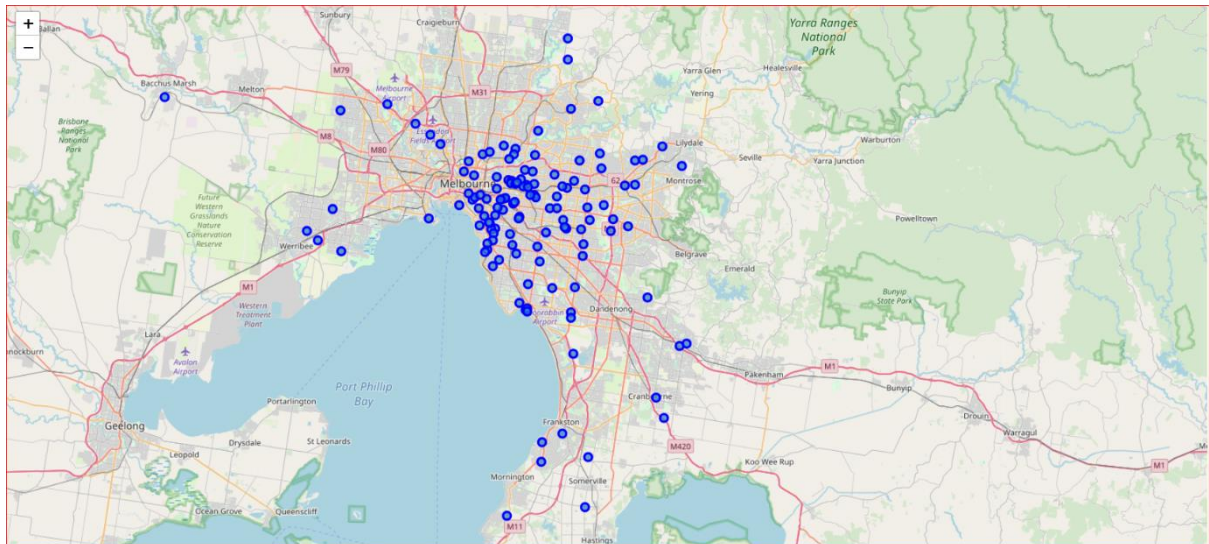|	School	Lat	Lng
0	Mac.Robertson Girls' High School,Melbourne CBD...	-37.836729	144.971831
1	Melbourne High School,South Yarra,VIC,3141	-37.835388	144.995677
2	Nossal High School,Berwick,VIC,3806	-38.039219	145.336030
3	Suzanne Cory High School,Werribee,VIC,3030	-37.893567	144.700048
4	Presbyterian Ladies' College,Burwood,VIC,3125	-37.848946	145.107125

The next step was to merge both dataframes, the geocoded schools with the schools' dataset that was already completed with all the relevant information.

Merge of the Lat/Lng dataframe with the Schools dataframe

```
[15]: #schl_vic
      schl_vic = pd.merge(melb_schl,geo_schl)
      schl_vic.head(10)
```
Out[15]:

	School	Postcode	State Overall Score	Total Enrolments	Sector	Lat	Lng
0	Mac.Robertson Girls' High School,Melbourne CBD...	3004	100	954	Government	-37.836729	144.971831
1	Melbourne High School,South Yarra,VIC,3141	3141	100	1357	Government	-37.835388	144.995677
2	Nossal High School,Berwick,VIC,3806	3806	100	832	Government	-38.039219	145.336030
3	Suzanne Cory High School,Werribee,VIC,3030	3030	100	871	Government	-37.893567	144.700048
4	Presbyterian Ladies' College,Burwood,VIC,3125	3125	100	1407	Non-government	-37.848946	145.107125
5	Fintona Girls' School,Balwyn,VIC,3103	3103	100	450	Non-government	-37.814764	145.080433
6	Haileybury College,Keysborough	3173	99	3927	Non-government	-37.992809	145.144701
7	Huntingtower School,Mount Waverley,VIC,3149	3149	99	697	Non-government	-37.876487	145.136121
8	Korowa Anglican Girls' School,Glen Iris,VIC,3146	3146	99	670	Non-government	-37.860997	145.054004
9	Camberwell Grammar School,Canterbury,VIC,3126	3126	99	1296	Non-government	-37.816624	145.066725

Given that the dataset has already the coordinates, the next step is to use the Foursquares API to retrieve surrounding venues in a rage of 1Km with a maximum of a hundred venues per school.

Methodology

With the geo-location of each school, the program will request to Foursquare API all the venues around the school. But before that, we can have taste of how is distributed the top secondary in the greater Melbourne in a folium map. This

Another important dataset that is required, are the venues around each school, for scope of this project we use the Foursquares API, that uses the Lat/Lng coordinates, max number of venues to pull out and the radius in meters around the given coordinates.

```
Out[29]: 'https://api.foursquare.com/v2/venues/explore?&client_id=I1MJ32QYFTODVVPW54MTDO4KK4
         BKTF4AZL4TSI5YU441UCMO1IMN4YHUAPRNQSNMBUY&v=20180605&ll=-37.8367287,144.97183090000

In [30]: results = requests.get(url).json()
         results

Out[30]: {'meta': {'code': 200, 'requestId': '5d76f852ad1789002c9beea3'},
          'response': {'suggestedFilters': {'header': 'Tap to show:',
           'filters': [{'name': 'Open now', 'key': 'openNow'}]},
          'headerLocation': 'Current map view',
          'headerFullLocation': 'Current map view',
          'headerLocationGranularity': 'unknown',
          'totalResults': 104,
          'suggestedBounds': {'ne': {'lat': -37.82772869099999,
            'lng': 144.9832054729414},
           'sw': {'lat': -37.845728709000014, 'lng': 144.96045632705864}},
          'groups': [{'type': 'Recommended Places',
            'name': 'recommended',
            'items': [{'reasons': {'count': 0,
               'items': [{'summary': 'This spot is popular',
                 'type': 'general',
                 'reasonName': 'globalInteractionReason'}]},
              'venue': {'id': '576ca21c498eecbf1b01fab5',
               'name': 'Crux & Co',
               'location': {'address': 'G01, 35 Albert Road',
                'lat': -37.834504818375116
```

The fetched information is .json file format, with the help of the previous labs, it was borrow the code to pass the information from .json to a dataframe as the below one.

```
print(schl_venues.shape)
schl_venues.head()
```

(3649, 7)

	School	School Latitude	School Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Mac.Robertson Girls' High School,Melbourne CBD...	-37.836729	144.971831	Crux & Co	-37.834505	144.971562	Café
1	Mac.Robertson Girls' High School,Melbourne CBD...	-37.836729	144.971831	The Kettle Black	-37.833992	144.971451	Café
2	Mac.Robertson Girls' High School,Melbourne CBD...	-37.836729	144.971831	Slater St. Bench	-37.839004	144.976725	Café
3	Mac.Robertson Girls' High School,Melbourne CBD...	-37.836729	144.971831	Kebabs R' Yummy	-37.831966	144.969831	Middle Eastern Restaurant
4	Mac.Robertson Girls' High School,Melbourne CBD...	-37.836729	144.971831	Tempura Hajime	-37.832900	144.968930	Japanese Restaurant

At that point, the data will be prepared to be fed in the K-means clustering to address similarities among the neighborhoods in Melbourne based on the top secondary schools. It is useful to remember that K-means works measuring how far is a school from the cluster centroid as the distance to the centroid decrease compare to other cluster's centroids, the sample (school) will assigned to the cluster where its distance is less from the cluster's centroid.

There are 281 uniques categories.

	School	School Latitude	School Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Mac.Robertson Girls' High School,Melbourne CBD...	-37.836729	144.971831	Crux & Co	-37.834505	144.971562	Café
1	Mac.Robertson Girls' High School,Melbourne CBD...	-37.836729	144.971831	The Kettle Black	-37.833992	144.971451	Café
2	Mac.Robertson Girls' High School,Melbourne CBD...	-37.836729	144.971831	Slater St. Bench	-37.839004	144.976725	Café
3	Mac.Robertson Girls' High School,Melbourne CBD...	-37.836729	144.971831	Kebabs R' Yummy	-37.831966	144.969831	Middle Eastern Restaurant
4	Mac.Robertson Girls' High School,Melbourne CBD...	-37.836729	144.971831	Tempura Hajime	-37.832900	144.968930	Japanese Restaurant

The dataframe consist of 281 different venue categories related to the 129 top secondary schools in Melbourne.

As the K-means algorithm does not recognize venues categories in characters, it is required to adapt the dataset to be numerical values. To do this we use the method get_dummies from pandas library to get numerical values per type of venues. Then this information is group by the School column to get the mean of the frequencies for each type of venues per school. The last step is to add the State Overall Score to this dataset scaled with only values from .9 to 1.0 as the top schools are ranked from 90 to 100 points.

|	School	Zoo Exhibit	Afghan Restaurant	African Restaurant	Antique Shop	Aquarium	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Aus Res
0	Academy of Mary Immaculate,Fitzroy,VIC,3065	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.01	0.0	0.01	0.010000	0.000	0.
1	Adass Israel School,Elsternwick,VIC,3185	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.0	0.00	0.000000	0.000	0.
2	Alamanda K-9 College,Point Cook,VIC,3030	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.0	0.00	0.000000	0.125	0.
3	Albert Park College,Albert Park,VIC,3206	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.0	0.00	0.033708	0.000	0.
4	Alphington Grammar School,Alphington,VIC,3078	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.0	0.00	0.000000	0.000	0.

At this point we have all the information that is required as input for the K-means algorithm. As explained before, we entry the mean of the venue types and the school score to the K-mean. The results from K-means is a list of labels for each one of the school entries as shown below in the "Cluster Labels" columns. The below table is a merge of the most common venues with the schools dataset already prepared with postcode, State Overall Score and Total Enrolments, Sector and Lat/Lng coordinates.

	School	Postcode	State Overall Score	Total Enrolments	Sector	Lat	Lng	Cluster Labels	1st Most Common Venue	2nd Most Common Venue
0	Mac.Robertson Girls' High School,Melbourne CBD...	3004	100	954	Government	-37.836729	144.971831	2	Café	Australian Restaurant
1	Melbourne High School,South Yarra,VIC,3141	3141	100	1357	Government	-37.835388	144.995677	2	Café	Italian Restaurant
2	Nossal High School,Berwick,VIC,3806	3806	100	832	Government	-38.039219	145.336030	2	Recreation Center	Tennis Court
3	Suzanne Cory High School,Werribee,VIC,3030	3030	100	871	Government	-37.893567	144.700048	1	Athletics & Sports	Health & Beauty Service
4	Presbyterian Ladies' College,Burwood,VIC,3125	3125	100	1407	Non-government	-37.848946	145.107125	2	Café	Gym

Results

Now we can put all that information on a map to visualize how the schools are clustered in the greater Melbourne area and answer the question of what are the similarities among the schools/suburbs of each cluster based on the features that the K-means had as input.

The cluster labels starts from 0 to 3, but for simplicity, we will refer from 1 to 4. So, cluster label 0 will be 1 and 1 will be 2 and so on.
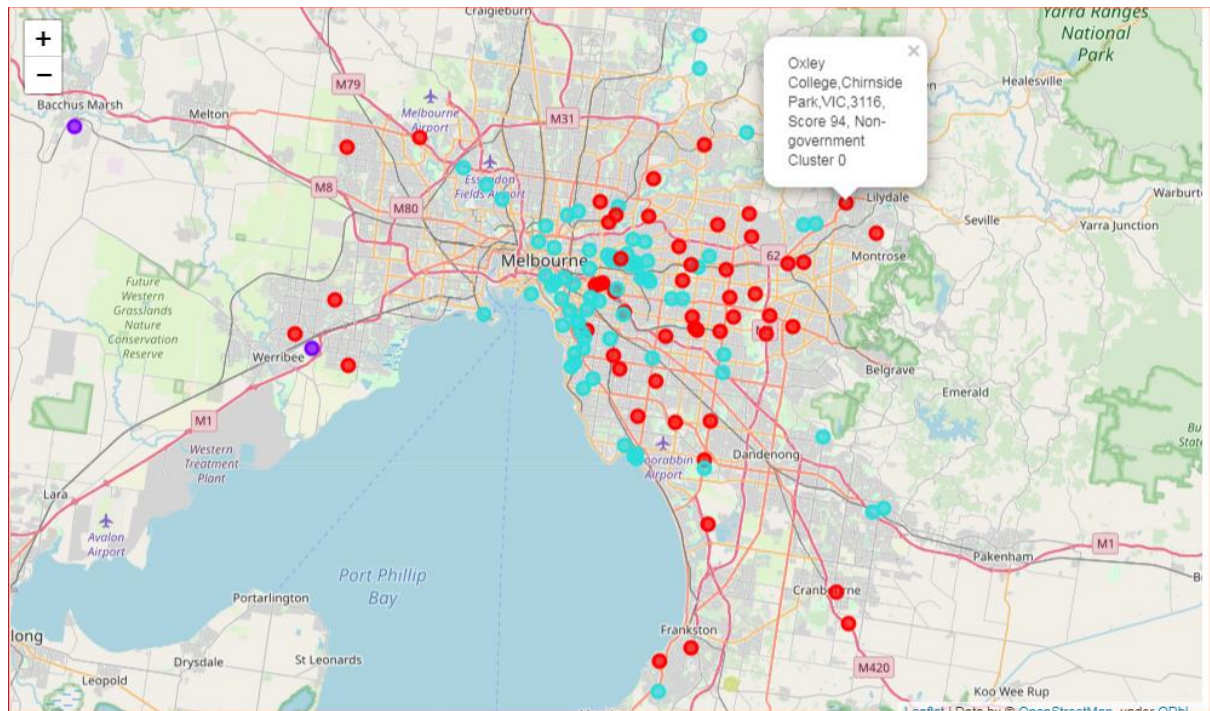
At first sight is observed 4 clusters, where most of the schools are group in 2 major clusters, one near the city's central business district or CBD and other in the outer suburbs of the city of Melbourne. One particular insight that is observed is that there are areas of low density or inexistence of top schools. At the west part of the city between Werribee and the CBD, limited at the north by the highway to the Melbourne's airport at the north-west. Then the outer north of the city there is a massive lack of top schools. The last zone that is appreciated is around the Dandenong municipality area, extends from the north in Montrose to the south, skirting on Frankston, and Cranbourne to the southern east through the Pakenham.

Note that K-means did not received as input any geospatial coordinates that allowed it to cluster by location of the schools, however, we can see that we have a correlation between the location of the school and the cluster assigned to the schools in the majority of the samples.

For clusters analysis, schools versus venues were grouped by the most common venue type as per below table. The analysis will done based on the two main clusters, cluster 1 and cluster 3.

Some fast facts about the school data.
- 129 rows in the data represents the Top schools in Melbourne by the 2018 year.
- 37 Government Schools are in top schools.
- 92 Non-Government School made it to the top.

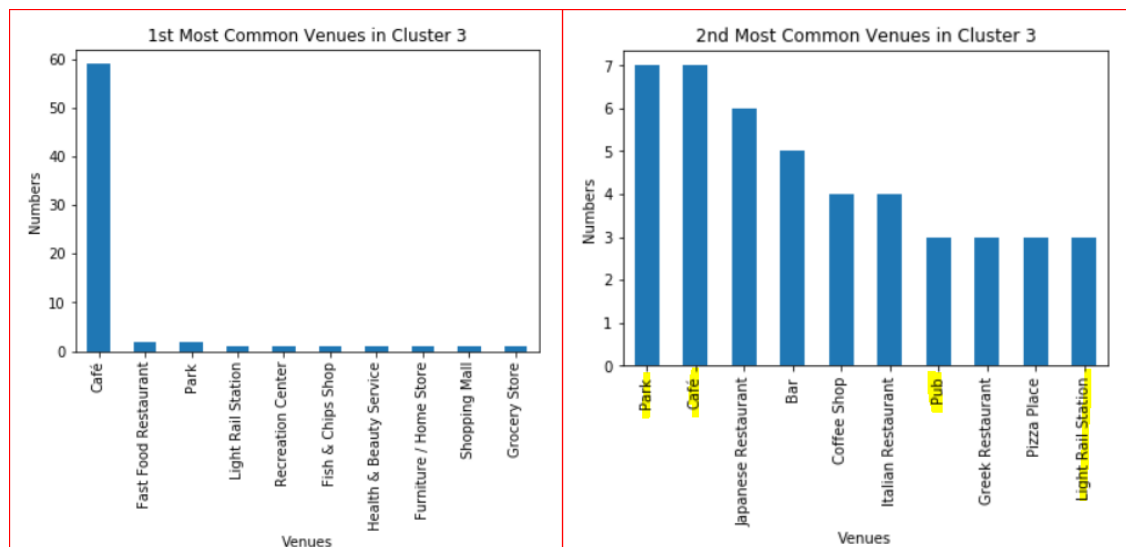|	Postcode	State Overall Score	Total Enrolments	Sector	Lat	Lng	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
6	3173	99	3927	Non-government	-37.992809	145.144701	0	Golf Course	Stadium	Ice Cream Shop	Greek Restaurant	Convenience Store	Electronics Store
7	3149	99	697	Non-government	-37.876487	145.136121	0	Pizza Place	Café	Japanese Restaurant	Australian Restaurant	Supermarket	Chinese Restaurant
8	3146	99	670	Non-government	-37.860997	145.054004	0	Park	Café	Vietnamese Restaurant	Train Station	Light Rail Station	Bakery
14	3142	99	2065	Non-government	-37.836002	145.025030	0	Park	Train Station	Supermarket	Athletics & Sports	Tennis Stadium	Baseball Field
15	3122	99	1866	Non-government	-37.834231	145.029162	0	Light Rail Station	Park	Train Station	Burger Joint	Athletics & Sports	Fish & Chips Shop

On the next two bar charts, it is presented the 1st and 2nd most common venues distribution in the cluster 1.
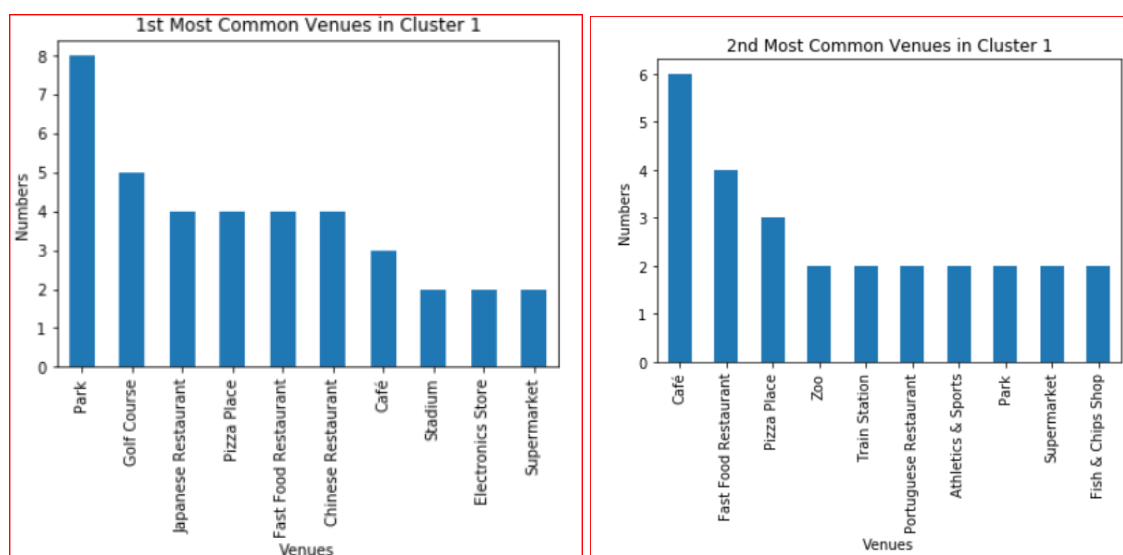
The first cluster is formed by 54 schools 14 funded by the government and 40 belong to the non-government sector.

Most common venues are parks, Golf courses, restaurants, cafes, etc. Seeing the map cluster 1 starts about 10Km away from the CDB of Melbourne and ends at rural part of Victoria State. This cluster of schools represents the typical suburbs out of the CBD, where families prefer to establish in order to be able to enjoy a more quite experience with family friendly amenities.

The 3rd cluster is formed by 72 schools. Most common venues are cafes, Fast food shops, parks, Light Rail stations (Trams),etc. Seeing the map cluster 3 is sitting of inner suburbs of Melbourne. This area of the city is much more dense in population as well in the supply of service and venues in the retail sector, one key metric is the amount of cafes that are close to the schools of this cluster, as well is the Tram station that are presents in practically all the inner suburbs of Melbourne.

Now we can have a look at the variables of the State Overall Score and Total Enrolments of Cluster 1 and Cluster 3. Note that State Overall Score was part of the feature given to the K-means cluster algorithm, on the other hand we will dig into the enrolment numbers to see density of students per cluster.

On the below charts we can appreciate that Cluster3 outperform Cluster1 as the median is 96 points over 95 in cluster 1. In cluster 3 the 75 % of the schools are above 93 points. Non-government school in cluster 3 are less scattered in range of values as its score is compressed in the 94-98 score marks whilst in cluster 1 we see the range from 92 to 98. Government sector in cluster 1, 50% of the schools are above 93 compare with 94 in cluster 3.


VICTOR ALVARADO P.
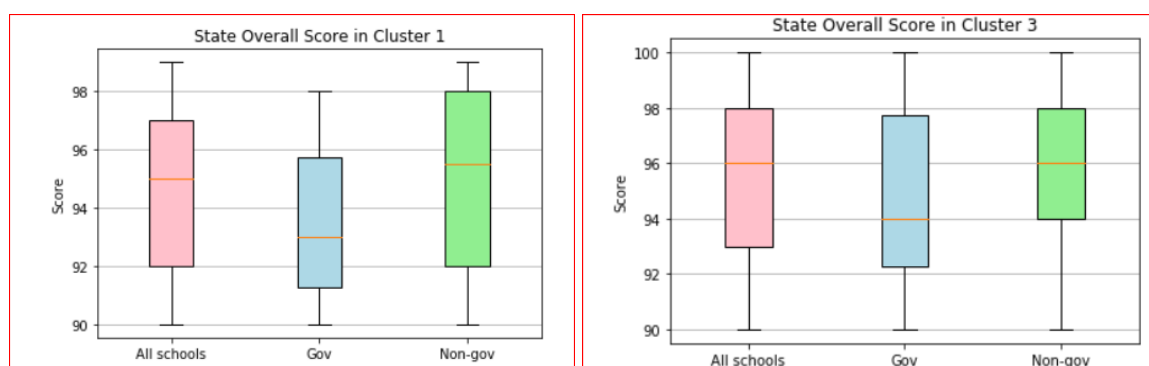
Cluster 1 and 3 State Overall Score tables.

	Gov Sector	Non-Gov Sector	Total
count	14.000000	40.000000	54.000000
mean	93.357143	95.100000	94.648148
std	2.437121	3.002563	2.947064
min	90.000000	90.000000	90.000000
25%	91.250000	92.000000	92.000000
50%	93.000000	95.500000	95.000000
75%	95.750000	98.000000	97.000000
max	98.000000	99.000000	99.000000

	Gov Sector	Non-Gov Sector	Total
count	22.000000	50.000000	72.00000
mean	94.909091	95.760000	95.50000
std	3.053633	2.825251	2.90216
min	90.000000	90.000000	90.00000
25%	92.250000	94.000000	93.00000
50%	94.000000	96.000000	96.00000
75%	97.750000	98.000000	98.00000
max	100.000000	100.000000	100.00000

Next it will presented the enrolment data for cluster 1 and cluster 3.

In Cluster 1 the government schools is observed more density in the enrolments with a mean of 1647.64 compare with 1070.9 of the non-government ones. The percentile are also showing more crowded schools in the public sector, where at least the half of the schools have around 1593.5 students compare with only 842 from non-government peers. From the total enrolment of the cluster 1, non-government institutions have almost 65% of the sample compare with the 35 % of the government's schools.

Total Enrolment in Cluster 1

Total Enrolment in Cluster 3
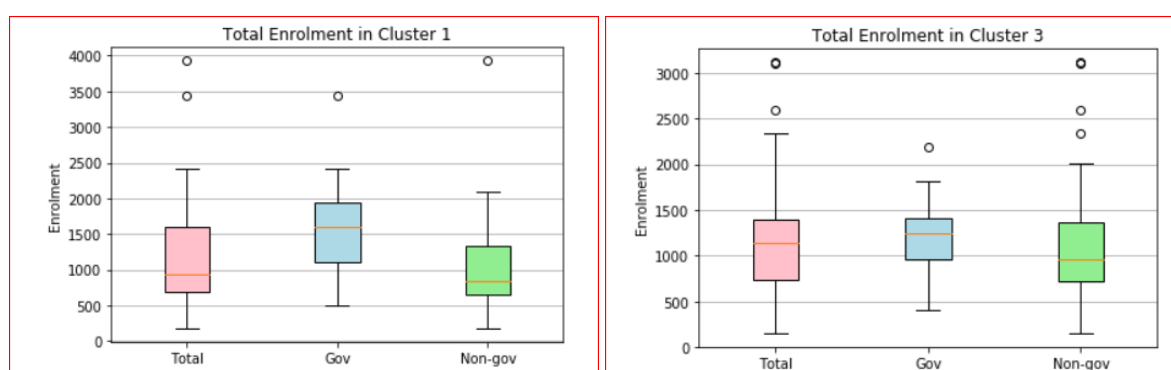
	Total Enrolments	Gov Sector	Non-Gov Sector
count	54.000000	14.000000	40.000000
mean	1220.425926	1647.642857	1070.900000
std	738.171443	737.241491	686.204716
min	180.000000	508.000000	180.000000
25%	688.750000	1102.500000	661.000000
50%	945.500000	1593.500000	842.000000
75%	1605.750000	1935.000000	1337.500000
max	3927.000000	3445.000000	3927.000000

Cluster 1 Enrolments

In Cluster 3 the government schools is observed almost same density in the enrolments with a mean of 1200.409 compare with 1137.82 of the non-government ones. The percentile are also showing more crowded schools in the public sector, where at least the half of the schools have around 1242.5 students compare with only 964.5 from non-government peers. From the total enrolment of the cluster 3, non-government institutions have almost 68.3% of the sample compare with the 31.7% of the government's schools.

The schools in this cluster are denser geographically with 68.3% of top schools.

	Total Enrolments	Gov Sector	Non-Gov Sector
count	72.000000	22.000000	50.000000
mean	1156.944444	1200.409091	1137.820000
std	576.190596	408.570290	638.977426
min	154.000000	406.000000	154.000000
25%	743.000000	963.750000	721.250000
50%	1141.000000	1242.500000	964.500000
75%	1398.000000	1417.000000	1370.250000
max	3114.000000	2183.000000	3114.000000

Cluster 3 Enrolments

Discussion

In the following points it will stated two observation and recommendations based on the information extracted from datasets.

- The geolocation distribution of the top secondary schools in Melbourne presents very notorious Blanc spaces that reflects an education divide in some areas of the city where there is no presence of top schools. In this regard, it has been identified areas of possible improvement for the Victoria Government in terms of the education.
- Cluster 1 that represents the outer suburbs of Melbourne, the performance of the schools in the State Overall Scores is lower than its counterpart in cluster 3, one part of the factor is that government sector drives the weight downwards in this feature of the data in cluster 1. In this scenario there is room for improvement in the government sector to uplift the quality of the education system.

Conclusion

After the application of cluster method using the K-mean algorithm, it was observed that two major clusters of secondary schools where gathered based on the similarities in terms of the venues/services available and the State Overall Score, that is compound metric of the results of each school in numeracy and literacy. The cluster 1 in red spots it is grouping schools based in suburbs that are away of the city CBD around 10km until the rural areas of Victoria. The most common venues are parks, Golf courses, restaurants, cafes, etc. This cluster of schools represents the typical suburbs out of the CBD, where families prefer to establish to be able to enjoy a more quite experience with family friendly amenities. The other big cluster is the 3rd one that is placed on the CBD are and inner suburbs of Melbourne. Most common venues are cafes, fast food shops, parks, Light Rail stations (Trams), etc. This area of the city is much more dense in population as well in the supply of service and venues in the retail sector, one key metric is the amount of cafes that are surrounding the schools of this cluster, as well is the Tram station that are presents in practically all the inner suburbs of Melbourne.

K-means clustering has been able to classify four clusters based on the services/venues and the State Overall Score metric. The results highlight that Melbourne's top secondary schools are divided in two main clusters, one in the inner suburbs of the city and the in the outer suburbs. From the methodology and results we can extract that inner city cluster performs slightly better in one percentage point over the outer cluster. In terms of the education sectors, Government or non-government institutions, the last ones outperform by 2 percent in the overall score.