The battle of the Schools by Victor Alvarado P.

Applied Data Science Capstone IBM/Coursera

Introduction-Business Problem

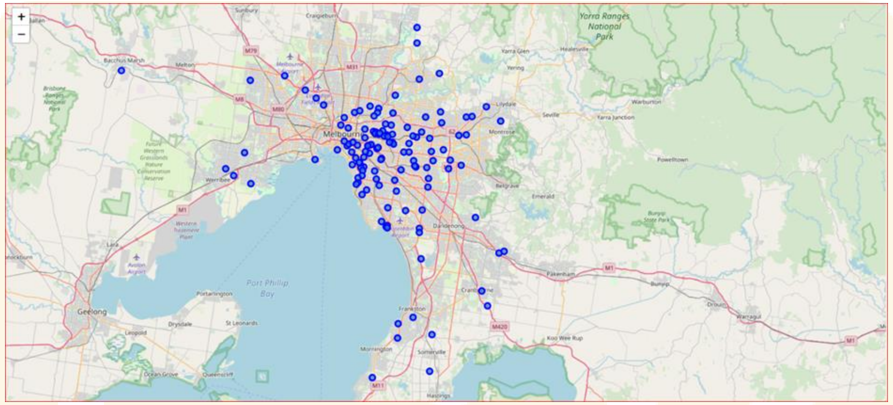
- In Australia the choice of a good secondary school is an important decision in families. Be government of private school, parents struggle in deciding which school to send their children, given the pros and cons in each sector of the education system. The project has the objective to produce an analysis to determine how group of schools can be clustered based on the suburbs/neighborhood where they are and their State Overall Score. This offers a tool in the decision process of choosing a school.

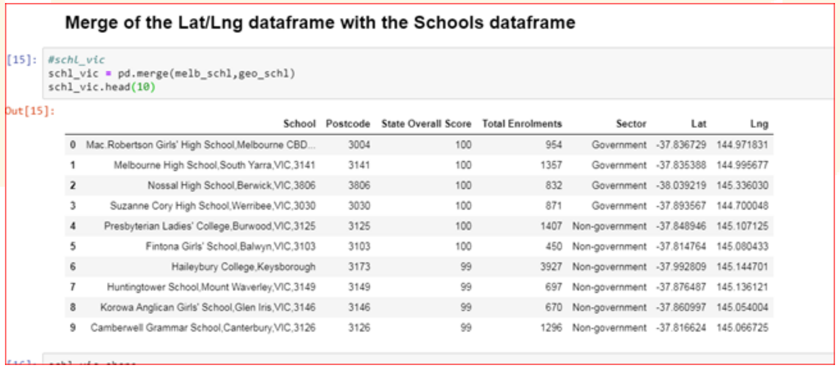
Data

- Our data comes from three sources:
- https://bettereducation.com.au/school/secondary/vic/vic_top_secondary_schools.aspx where we scrapped a table with the top secondary schools in Victoria state. The features we used in this project were the school name, State Overall Score and Total Enrolments. In the School information there is also the postcode and the suburb name where the school is. In many of government sector the information the Suburb was missing and thus needed to be obtained from another sources.
- The second source was https://en.wikipedia.org/wiki/List_of_Melbourne_suburbs where we got the complete suburbs and postcode information for the Victoria state. Along with this we use the geocode add-on of Google Sheets to process the geospatial coordinates.
- The third is the Foursquare's API that will provide us with the surrounding venues of the schools in order to give us a suburb profiles of each schools in out dataset.

Methodology

- After the information in the datasets were cleaned and filled with missing information from our two data sources.
- A dataframe with School metrics and geospatial coordinates was built, allowing in its first instance to visualize the location of each of the top schools in greater Melbourne.
- The first insight observed was that the are areas of Melbourne with very low density of top secondary schools.





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)
```

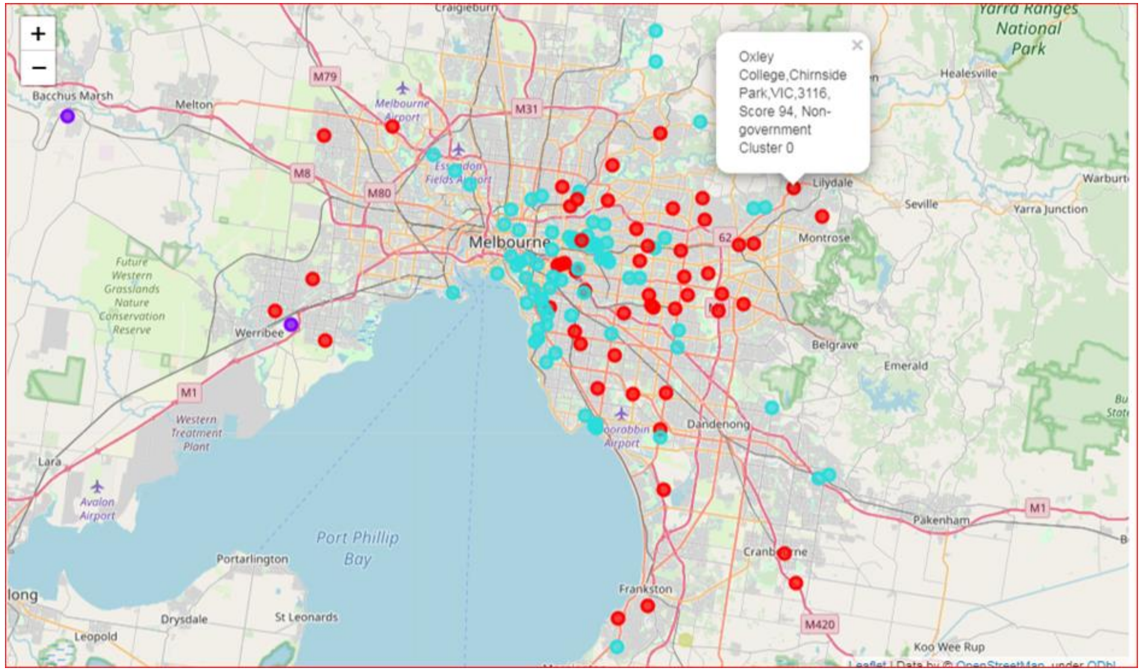
Out[15]:

	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

Methodology

- The most important element of the process was to prepare the information required by the Machine Learning algorithm (K-means) to produce a result, in this case clustering schools based on the State Overall Score and their suburbs(neighborhood) profile.
- This was represented by the information that came from the Foursquare API, that fetched the venues around with a radius of 1km.
- With the venue’s category feature, that says if there venues was an café, Italian restaurant, etc., that relevant information was transform in to “dummy values” of “1” or “0” were there is no venue of that type in the suburb of a sample of the data. With the previous method the dataset’s samples as many venues categories were found. Then a new dataframe was created grouped by school with the mean of each venues frequency.
- Adding a scaled State Overall Score to each entry, the data set was given as an input to the K-means algorithm to process K = 4.

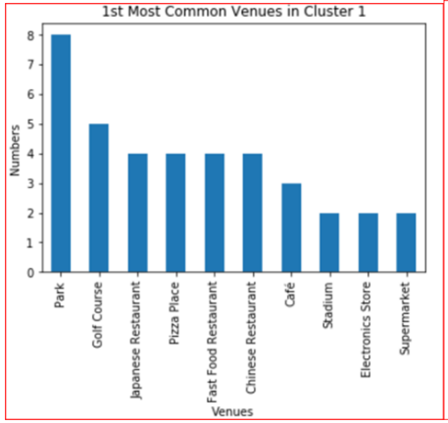
Methodology

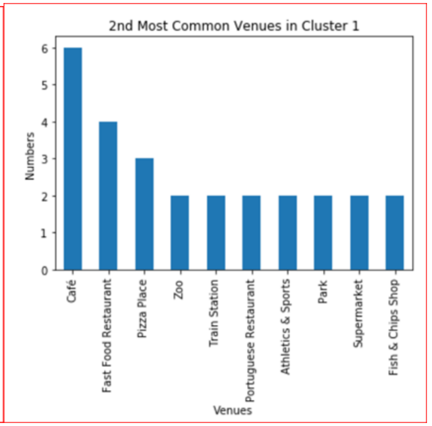


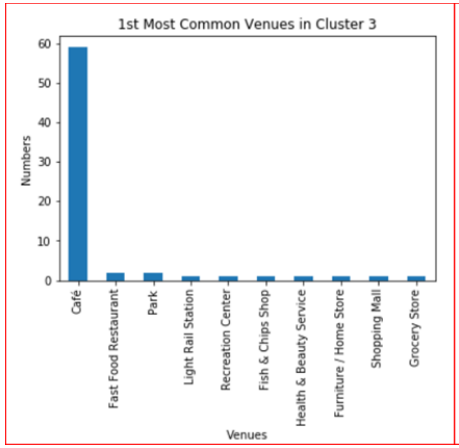
- The clustered dataset of K= 4, were formed in two major clusters.
- As two of the cluster were simply of 1 and 2 schools, they were not analyzed, as they are not a representative sample of the data.

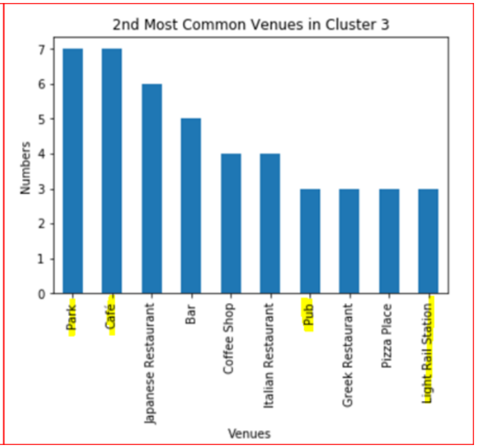
Results / Discussion

- After we obtained the cluster labels for each school, the next steps is analyze the clusters by the most common venue categories in each cluster. For simplicity we present the 1st and 2nd most common.
- With the aid of the bar char we can identify if the cluster is in inner Melbourne or in the outer suburbs. Cluster 1 counts more parks, golf course, restaurants and cafes, that matches with the outer suburbs, on the other hand cluster 3 has a high frequency of cafes and also we see trams station as 2nd most common, this represents the inner suburbs.



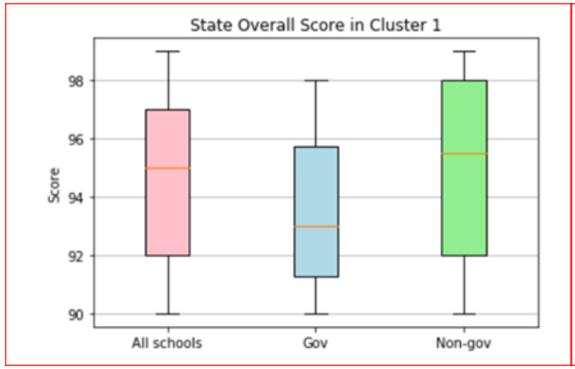


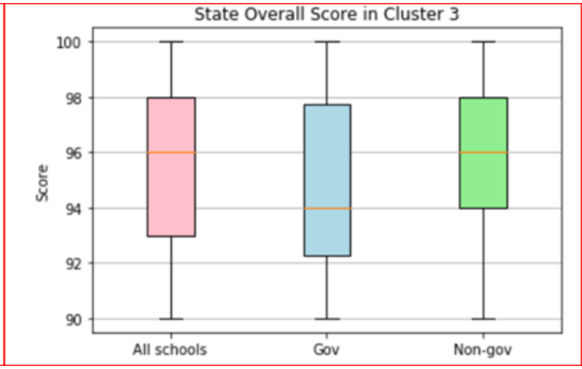


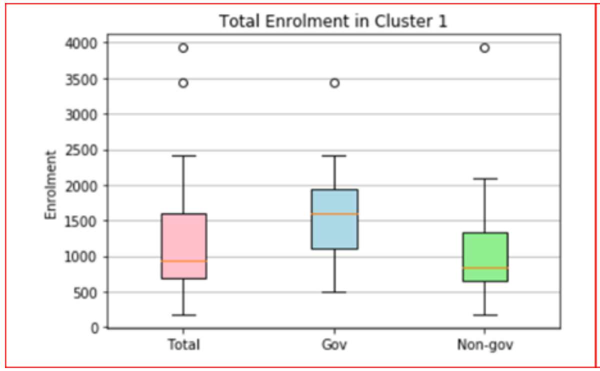


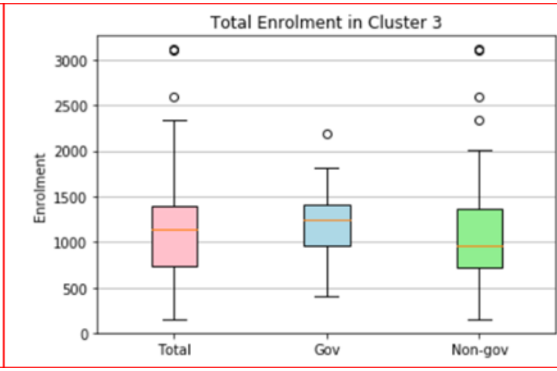
Results / Discussion

- 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 its score is compressed in the 94-98 score marks whilst in cluster 1 we see the range from 92 to 98.
- The student density in cluster 1 is bigger that in cluster 3 and is driven by government sector, despite there are some outliers in both sectors. Non-government schools in cluster 1 is less than in cluster 3, which is correlated to the inner suburbs where there is more population.









Results / Discussion

- 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.

Conclusions

- 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..

Applied Data Science Capstone IBM/Coursera

The Battle of the Schools by Victor Alvarado