*Melbourne Education Facility Distribution*

**2) *Domain***: Education

**3) *Question***:

* Are education facilities reasonably distributed across Melbourne?

**4) *Datasets***:

* There are 3 datasets that are involved in this project. All of which, have been updated and replaced several times throughout the project.

***Details***:

* The main dataset is “*allschoolslist2016.csv*”. It contains detail information about all the schools in Melbourne; such as name, location and specific coordinates. The quantity and quality in this latest version of school list data have largely improved from those that I previously obtained.

School list data link: <https://www.data.vic.gov.au/data/dataset/school-locations-2016>

* Next dataset is “*population\_parse.json*”. This dataset consists of population data for most suburb in Melbourne. This dataset is actually obtained by myself, using *web scraping* method and collect information online. The reason of this is due to the lack of sizeable population datasets available for download.
* Also, since I’m processing education facilities. Having only school dataset seems to be quite limiting. Hence, the decision to bring in one more dataset consisting library information, “*libraries.csv*”.

Library: <https://www.data.vic.gov.au/data/dataset/libraries>

**5) *Pre-processing***:

* The workflow of data pre-processing alternates with different data formats and each update of datasets. The usual processes I went through for data processing are as follows.

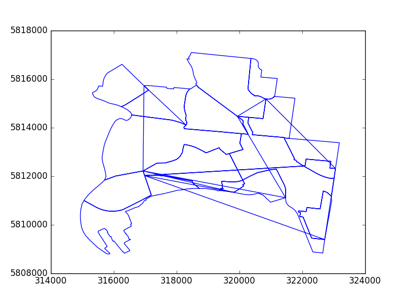
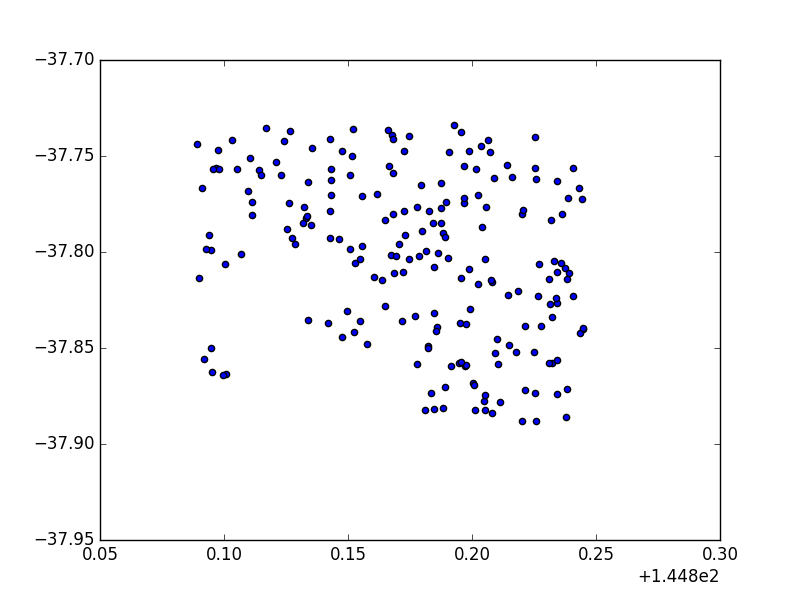
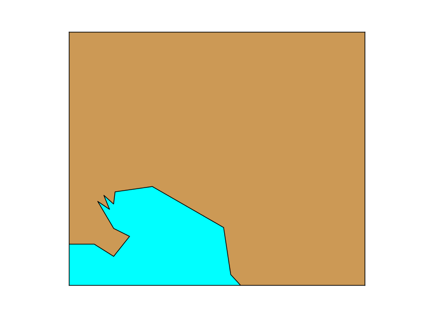
***Diagnosing data problems***:

* Initial planning for any datasets are critical. Before making decision to select any dataset, the immediate step will be diagnosing for potential issue in the data. I remain sceptical even to some promising datasets, until I have diagnosed the data. Since each dataset is unique and comprise of its own bugs or problems. Having recognise those factors would reduce future errors and greatly improve efficiency.

***Visualizing raw Data***:

* Once a dataset has been determined after the diagnosing processes, a trial visual assessment will be implemented on the raw data. The visual evaluation is to analysis for usability and specific methods for subsequent visualisation.

Here are some examples of my initial visuals from raw data.



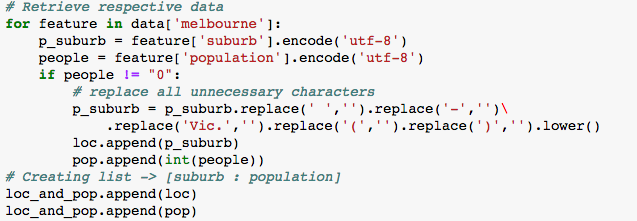
***Data Cleaning***:

* At this point, I would be confident enough to commit to the dataset. Thus, the beginning of data cleaning process. The objective here is resolve and fix to the greatest degree, for those previously diagnosed problem within the dataset. This includes correcting erroneous values and excluding redundant information.

***Transforming Data***:

* Although the data has been cleaned, it is nonetheless not particularly useful. The dataset is merely set for the transformation process. In order to transform seemingly random strings to feasible useful data, there are various modification it needs to undertake. Generally, those procedures are data formatting, extraction, refinement, and conversion.

Below is a sample from the codes I have written for data transformation. In this case, my data is in ‘Unicode’ format, which needs to corrected to ‘utf-8’ as it causes unwanted list content. Also, replacement function is used to increase data compatibility.



***Limitation***:

* My data pre-processing methods does appear to be promising. However, there a major limitation, which is time. When time is a concern, my pre-processing does take up considerable amount of time to finally prepared my dataset. This is especially true as my question alters throughout the project, leading to several major update of datasets.

**6) *Integration***:

* Data integration is the final stage of data wrangling before visualisation. The result generated hereby have direct impact on the heavy dependent visualisation. Therefore, confirmation of data accuracy is the priority.

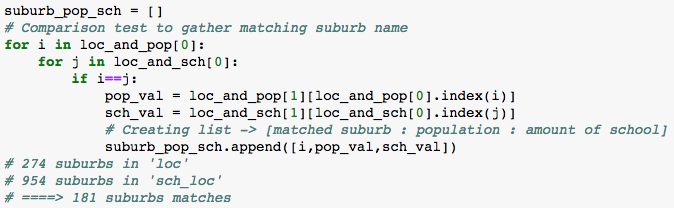
***Data comparison***:

* As part of the project requirement, integrating multiple data sets is unavoidable for combining datasets from different sources into meaningful and valuable information. Hence, data compatibility is an important factor to consider. Previous dataset selections, planning and transformations are all preparation for this moment. In those datasets I have selected, the comparison factor is the location, specifically the corresponding suburb of each school, library and resident population. In order to proceed with data comparison, I have to pre-construct various lists and arrays with a consistent location string valve from respective datasets.

***Matching and combine datasets***:

* After completing the preparation of data comparison, the matching and combining of clean implemented lists or array are much peaceful to process. Simple loop corresponding matching element to extract similar data index, then append those combination of matched component into a single data list.

In the example from my code shown below, two formulated data list are being compared and matched by indexing similarities. Thereby, combined together to create a new data list that can be further processed for visualisation.



***Limitation***:

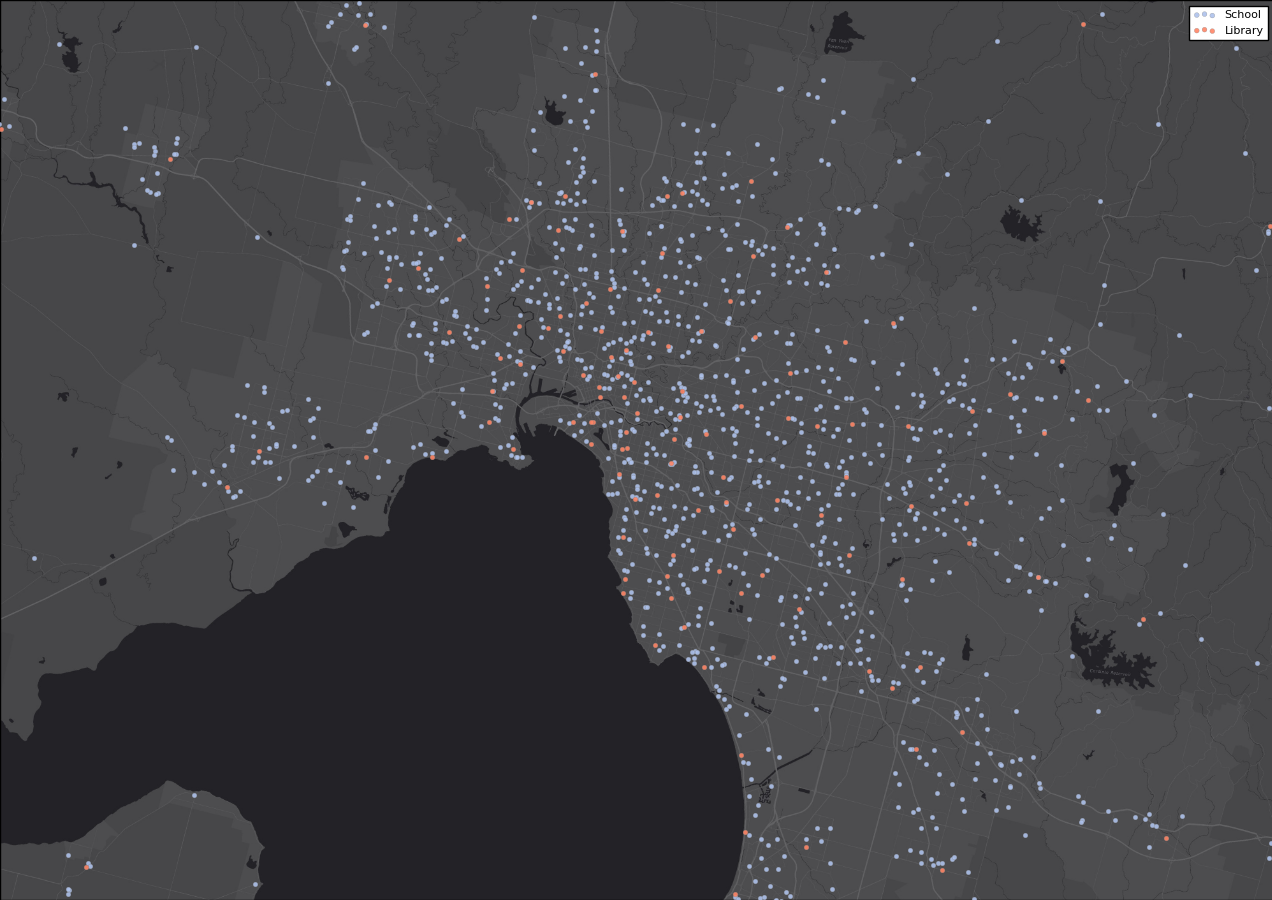
* The method of my data integration does worked to a certain extend. As shown on the example above, resulting data size has largely been reduced in proportion after combination compare to before. This indicates that the choice of datasets does play an important role in this case. In a scenario of my earlier datasets, where raw data samples are very limiting in numbers. This could cause major issue, as the possibility of matching both datasets yet retaining conclusive outcome are improbable.

**7) *Results***:

* I have created a total of 4 distinct visualisations with different analytical significances. Proving investigation from a broader perceptive to a narrower range targeting specific outliners.

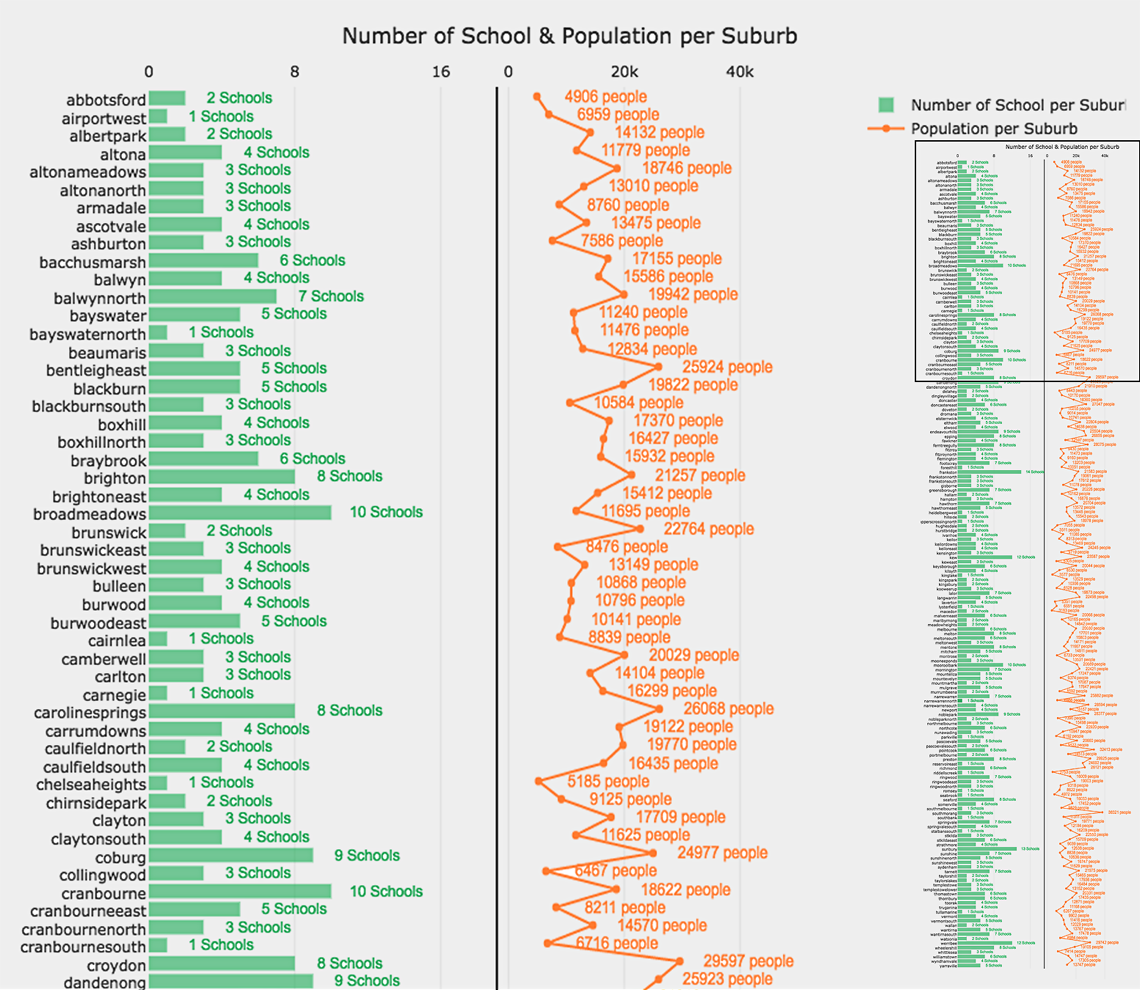
***Map visualisation***:

* This is a satellite view of the city of Melbourne, with indication of educational facilities scattered through the map. As legend implied on the visual, school and libraries are pin pointed on its exact location. This map would effectively introduce and present the general impression of the area of clusters and outliner. Thus opening the sense of entry to further detail visual analysis.



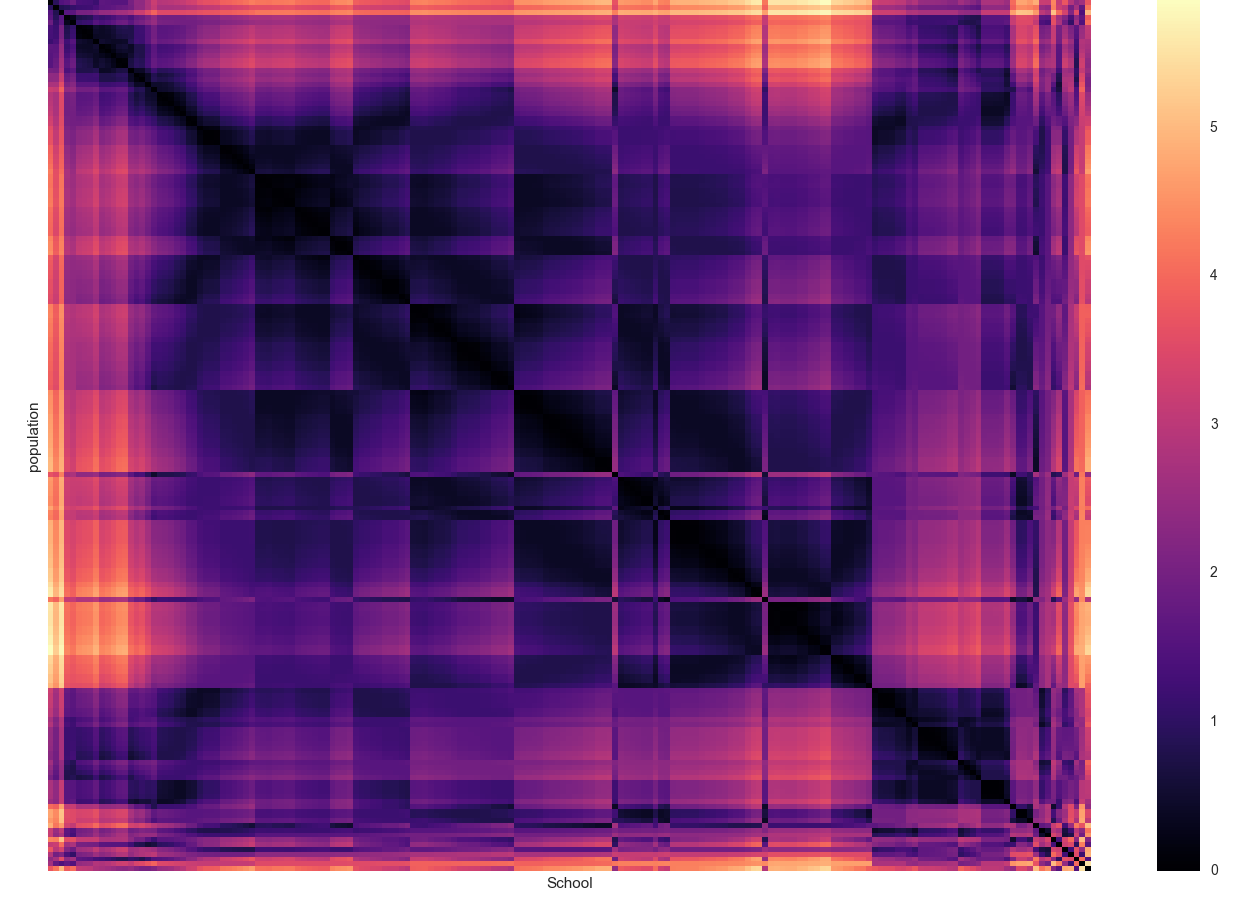
***Comprarison summary***:

* This visualisation provides a comprehensive list of detail information for each individual suburb in Melbourne. In every matching suburb, there are comparisons of school numbers and the residential population. The adventage of obtaining this visual not only allow searching specific targeted information possible, but also opens other potential analytical values. Since the conclusive visual could easily be tranformable and derived into various relevent information.



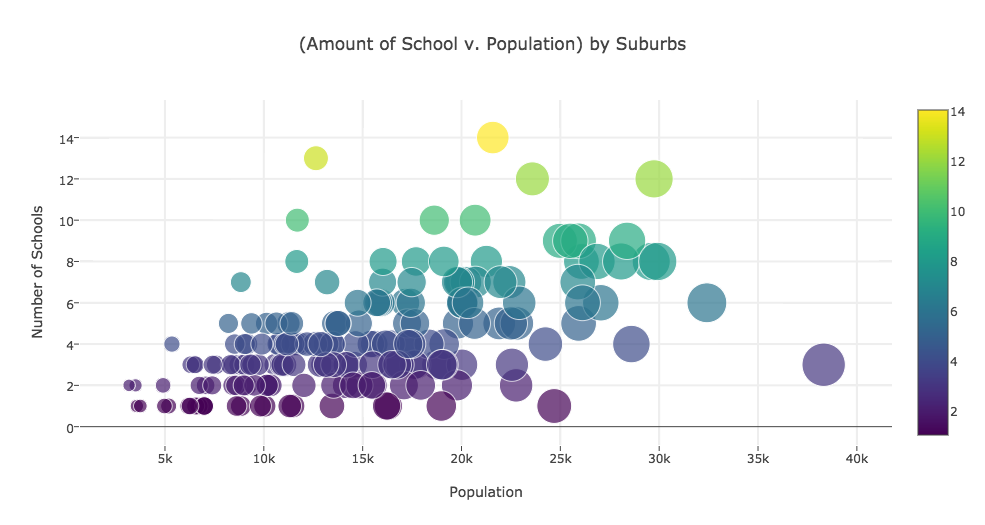
***VAT algorithm***:

* The motivation of having this visual is to locate general clusters and trendings. In this case, the VAT visual of the comparsion dataset displayed 2 different clusters. One of which is significantly larger than the other. This implicit an idea of general clustering, suggesting that most schools are evenly distributed throughout Melbourne except for a small amount of outliners. Therefore, focusing attention on specific problematic ares.

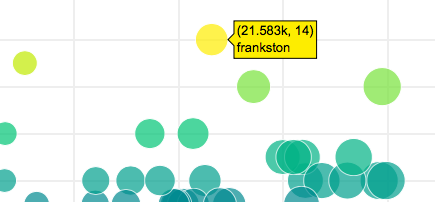


***Bubble Overview***:

* Here is the last visualisation of the project. It demonstrates a clear visual conclusion from general trending to individual targeted outliner. Thereby, accomplished my final data analysis with clear presentation of distict trends and outliners.



* Adding to the complexity, this visualisation generated has enable interativity. Hence, one could directly extract information of any particular data bubble. All 3 relevent information of population, number of school and location will be shown.



***Limitation***:

* Overall, my visualization process did succeed in presenting a clear conclusion of my question. They provide different visuals, representing clusters and trending, as well as direct separation of outliner. However, the limitation of my approach might be the lack of solution. Although, finding a solution to my question is not a specified part of the project. It would be much helpful to include such idea in my visualisations.

**8) *Values***:

* It will surely be a hassle, if anyone tries to make sense and interpret any useful information by looking at the raw data. After the data processing and integrating, much relations has formed between different datasets. At which, usable information has been abstracted in collation of locations, school numbers and resident populations.

The real value is added after the visualization. Mapping various information from datasets and visualise them have allowed deeper understanding of information be able to be captured at a glance. This will save government valuable time for any future planning of education facility allocations.

**9) *Challenges and Reflection***:

* The challenges in pre-processing and integrating are certainly the data transformation and comparison. Both are crucial within their sector, and every stage after which are very much dependent on them. Substantial amount of time has been utilized on figuring out the methods of data wrangling.
* Visualisation post huge difficulties as I have envisioned at the beginning of project. Hours of research are done in search of creating precise and outstanding visualisation. At the beginning of raw data visualising, I have been testing and exploring various method to visualise my data. Throughout the project, I have to self learn my way across many python libraries that could potentially improve the resulting visualisation. Consistent effort has put into enhancing and optimise visualisations, for all visuals I have kept records of different versions and only present the best.
* Dozens of unsuccessful attempts have been encountered for very visualisation I have created. The most grinding and strenuous has to be the map visualisation, as the ‘Basemap’ library is extremely had to understand. Plotting with ‘plotly’ isn’t much comfortable as well, due to the amount of data I had to manipulate.

**10) *Question Resolution***:

* With this question answered, a clear picture of education distribution in Melbourne will be rendered. By comparing the number of residences and the availability of facilities in each location, we could comprehend for any unbalance differences. Thus, educational infrastructures and facilities could be re-evaluated and planned for future improvement. This will be favourable to parents and certainly beneficial to the students across Melbourne.

***Person of interest***:

* The minister of education and the Mayor of Melbourne city will be interested in the answer of this question. For education is such an important matter, that many local residences, especially parents will be concern in. Furthermore, an improvement in education sector has always been the priory of our government.

**11) *Code***:

* All my programming in this project is coded using Python and written on ‘Jupyter’. Initially, I imagined to write about 200 line of codes. However, I have underestimate the scale and the amount of work require to complete the project. Finally, I end up with around 800 line of codes in the process. Vast majority of which is written from scratch, for some parts are referenced from helpful website such as ‘stack overflow’.
* I have make used of numerous python libraries in throughout the project. While data processing and integrating are concern, ‘numpy’ is the most useful library. However, the most important library of all will be ‘Basemap’ , ‘matplotlib’ and ‘plotly’. My visualisation could not be achieved without them.
* Here are all the python libraries used:

