**ETC1010 project data description**

* **Group members**

|  |  |  |
| --- | --- | --- |
| Boqin Wang | bwan0023@student.monash.edu | 29646952 |
| Yuhan He | yhee0050@student.monash.edu | 30445299 |
| Bocong Zhao | bzha0020@student.monash.edu | 28305329 |
| Bingqing Hua | bhua0008@student.monash.edu | 29930847 |

* **Data description**
* **Data**
* VCE & ATAR Results 2018: The top government and non-government schools in Victoria, which comes from [The Age: Latest & Breaking News Melbourne, Victoria](https://www.theage.com.au/)

The URL is <https://www.theage.com.au/national/victoria/vce-and-atar-results-2018-the-top-government-and-non-government-schools-in-victoria-20181219-p50n65.html>

* Top 100 Ranking VIC Melbourne Secondary School in 2017 (Non-government and Private). The URL is <https://www.schoolcatchment.com.au/?p=10044>
* Top 100 Ranking VIC Melbourne Secondary School in 2017 (government)

The URL is <https://www.schoolcatchment.com.au/>

* **Data includes**
* The 2018 VCE perform for 580+ top Secondary School located in Victoria.
* Top 100 government and Private Secondary School located in Melbourne respectively.
* **Aim of project**

According to themedian VCE study score, percentage study scores 40 and more respectively to answer the following questions.

* Whether public and private secondary school influences students’ performance in their VCE sources.
* Whether the Secondary School located in different area in Melbourne would really influence students’ performance in their VCE sources.
* Is there any extreme situations, for example:
* there are the school ranking or median VCE study score is low, but there are many study scores 40 and more.
* there are the school ranking or median VCE study score is high, but there are fewer study scores 40 and more.
* **Selected variables**

The data has many parts, however, our research just want to see the name of school, locality, rank, median VCE study score, percentage study scores 40 and more, school sector, total enrolment etc. Thus, we have excluded other parts of data.

* **Process of exploring data/ Different forms of data and making them tidy**
* Use clean\_names() to make names unique consist only of the \_ character, numbers, and letters.
* Use select() to choose the variables to be observed and eliminate useless rows in the datasets
* Use rename() to label the variable with its actual meanings
* Use vis\_dat() and vis\_miss to figure out the missing value and use drop\_na() to eliminate them
* Use mutate() to add new variables
* Use “as.numeric” , “as.factor” to transfer the nature of variables, eg: numerical, category
* Make dataset as.tibble() then rbind() to joining those three datasets together in order to figure out the sector of school
* The dataset after tidy named “school\_tidy”
* Have a briefly view of the relationship between each variables in the dataset