Project 1 – Structure Plan

Preparation

Part1 – data cleaning

1. Cleaning all data, such as reduce size by removing irrelevant columns and NAN value rows, combining csv files to a simple appearance, changing column names for later merging, converting cell types etc.

Part2 – data checking

1. Calculating outliers of data to check if the data is good for use

Analysing

Part1 – attendance against school types and years

1. Attendance rate vs. years (2015-2019)
    a. Use the average attendance rate of all school listed (1228 schools) for different years
    b. Line plot
    c. Note: some schools don't have 5 years of data recorded, hence use the average value (attendance rate)
2. Attendance rate vs. school types
    a. Use average attendance rate on each of the school types, totally 4 types
    b. Bar plot
    c. Note: suggested to have the average value of all years, however can also have 5 bar plots for each of the year, should be depended on the number of data based on each year *require discussion

Part2 – attendance against different areas

1. Attendance rate vs. Region
    a. Use the average value of all school listed within the same Region, totally 7 different regions
    b. Pie plot
    c. Note: same concept as 2c in part1 *require discussion
2. Attendance rate vs. LGA
    a. Use average value of all school listed within the same LGA, totally 74
    b. Bar plot
    c. Note: there is an option to use a line plot method with the attendance rate as the y-axis, LGA as the x-axis and 5 lines representing for each year, the reason is that, 74 different LGA is a bit too much as an axis for bar plotting, and as there are missing data for schools in a typical year, mixing up them may providing misleading trends, whereas analysis against each year could generate more accurate outcomes *require discussion

Part3 – attendance against socio-economic score

1. The file of SEIFA is providing us with the state-wide decile ranking for each postcode, ranking level 1 to 10, with the higher the ranking, the better the socio-economic environment
2. There are totally 399 postcodes recorded from the schools and can be used to match up with the score in SEIFA file
3. Scatter plot
    a. Using scatter plotting with attendance rates of all school (average value of all years) against the socio-economic scores
    b. Applying line of equation and r-value to see the correlation
    c. Try to change the x-label as the 10 rankings instead of the scores to make it fancy

Part4 – attendance against LGA offence rates

1. We will match the same years of data along the LGA names to attendance rate file
2. Some of the offence may irrelevant to school attendance, may consider to remove the column
3. Scatter plot
4. Using average attendance rate of all years against the average count on all related offences (or all offences) of all years located in each LGA
5. Applying line of equation and r-value to see the correlation
6. Note: before plotting, ascending the number of offence counts and list it up as each LGA, then make a bin of at least 5 ranks (I will suggest 10). And same as 3c in part3, change the x-label to each of the ranks
7. There are only 74 different LGA and I am thinking to have something more to make the correlation feel more reasonable *require discussion

Part5 – Heatmap plot of Googlemaps API calls on school attendance rates

1. Introduce API calls as a bonus
2. Could have more interactions here, such as having markers of all police stations etc.

Reporting

Concludes all findings based on previous studies

Presentation

1. Discuss about how to present, i.e. only by a readme file or word documents or PPT etc.
2. Run presentation trails to check the time

Time Frame of project

1. Coding – including preparation and analysing, should be done by 23rd of June
2. Reporting and examination – by 25th of June
3. Presentation preparation – by 28th of June