5/21/2019

# COMP20008
# Data Science Project
# Group 106

Long Chen (1077055)
Mian Chen (941099)
Yuntong Jiang (940590)
Chris Xiu (1154943)

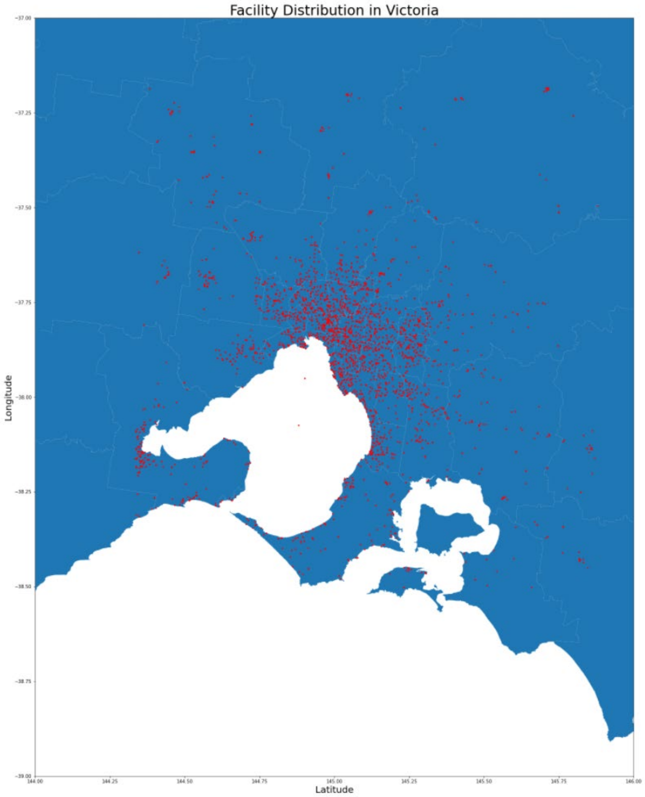
# 1. Introduction

Physical activity is crucial for a person's health and wellbeing. In order to play a sport or to exercise, people will usually go to a park, stadium or gym. Our team's research question is "Will the number of sports facilities affect the health of residents in different regions of Victoria?". This study investigates how the accessibility to sports facilities impacts people's willingness to exercise and their physical health within each area. The results of this research topic will help local governments to understand whether increasing the number of sport facilities can potentially improve the healthiness of the residents and the livability of the community.

# 2. Method

## 2.1. Original Data

The two main open datasets from Aurin that we are using are "Victoria Sport and Recreation Facility Locations 2015-2016" and "LGA15 Adults Health Risk Factor Estimates 2014-2015". The first dataset is a csv file containing approximately 80 types of sport facilities including private gyms and fitness centers, there are roughly 5000 of them in total across 40 local government areas (LGA) in Victoria. The plot below shows the distribution of these facilities plotted on a map.



The second dataset contains the estimated percentage of people aged 18 years and whose health was at risk in each of the following categories: psychologically distressed, high blood pressure, overweight, obesity, waist measurement, smoking, alcohol consumption and no or low exercise. Note that the two dataset aren't from the same year, however the number of sport facilities is unlikely to change much in one year so it won't affect the quality of our analysis.

Another two datasets we used are “Victoria Local Government Area ASGS Edition 2016” (ASGS 2016) and “Regional Population LGA Vic” both from Australian Bureau of Statistics. The first dataset is used to link the two major datasets together as explained later in the Data Wrangling part of this report. The second dataset provides us with relevant information about the population and area of each LGA, which is important when considering the number of sport facilities in proportion to the land size or population of that LGA. As simply using the nominal amount does not reflect the actual density of sport facilities in each area.

### 2.2.Data Wrangling

We performed a lot of prepossessing to all datasets, as they are fairly large and contain useless and even incorrect data. First, we extract the information about the LGA names and the corresponding LGA codes from ASGS 2016, then we use it to match the LGA name given in the Victoria Sport and Recreation Facility Location. For those locations we cannot find a LGA code to match, we remove the data. Next, we group the sport facility in the same area together by LGA code. Then, we join Victoria Sport and Recreation Facility Location, LGA15 Adults Health Risk Factor Estimates 2014-2015 and LGA regional population together by LGA code and get a new data frame named JOINED2. Furthermore, to normalize the data, we divide the number of facilities per the population and the area size respectively.

### 2.3.Data Analysis

We use both simple linear regression and log-scale quadratic regression to predict each of the health indicators, as we are not sure if the relationship is linear or not. Based on the output from the models, using measurements such as coefficient of determination(R^2) we can tell how much variation of the dependent variable is being explained.

Next, we decided to use decision trees to perform classification. When using simple linear regression, we are treating the health variables as continuous. However, it can also be considered binary, as the health status of the group of people who do not do any exercise at all and group of people who do. Hence, we want to see if the density of sport facilities is helpful to classify the health variables using decision trees.

Lastly, instead of focusing on the amount/density of sport facilities, we want to investigate which type of facilities are the most popular among the healthiest areas.

## 3. Results

### 3.1.Correlation

Firstly, based on data grouped by LGA, we computed Pearson correlations between the number of facilities per km^2 and the percentage of people suffering each type of risk.

| **Risk** | **Correlation** |
|---|---|
| Did low or no exercise | -0.66308798 |
| obesity | -0.85585843 |
| high blood pressure | 0.2094712 |
| psychological distress | -0.47607278 |
| current smokers | -0.70782562 |
| risky alcohol consumption | -0.06628064 |
| risky waist measurement | -0.81370185 |

Except for high blood pressure, all risk indicators are negatively correlated with facility density. The correlation against alcohol consumption is slight, while those against other risk factors are strong.

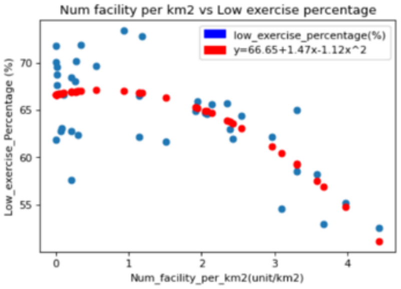
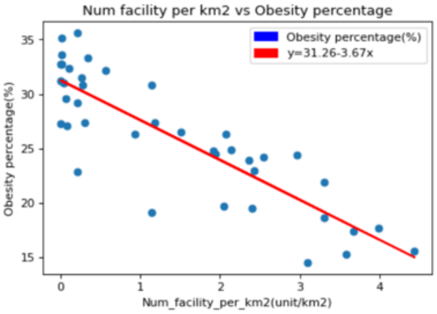
## 3.2.Regression

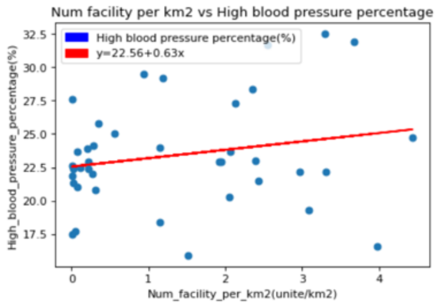
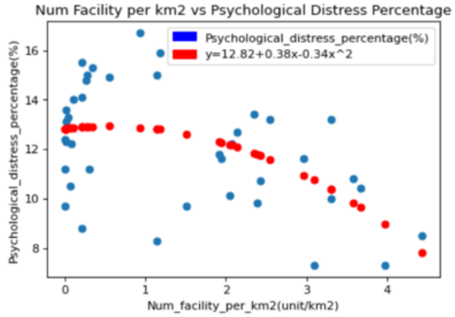
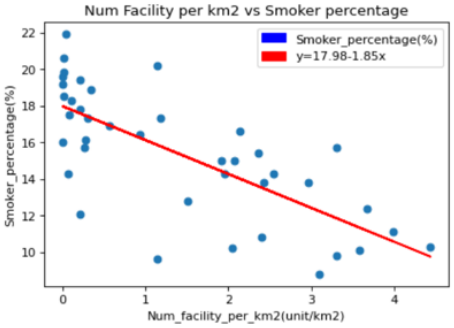
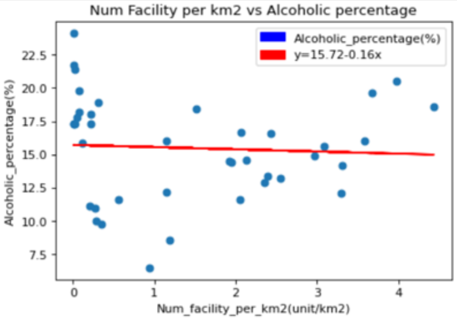
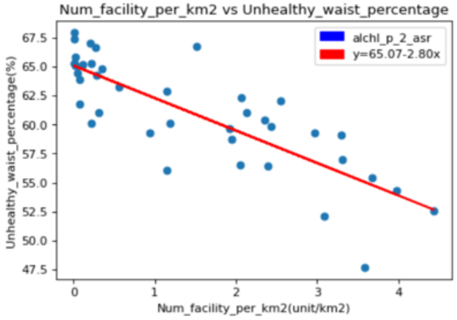
Regression analysis was conducted on the presence of each type of health risks (per 100 population) against the number of facilities per km^2. Each LGA contributes one observation.

To better explore the main relationship between variables, we firstly removed one high-leverage observation with facility_per_km2 of 6.82. Then, we performed simple linear regression, polynomial regression, and log-scale polynomial regression for each health risk indicator and selected the optimally fitting models.

Optimal models:

X_i: the sports facilities per capita per square kilometer

| $Y_i$: | Model | $R^2$ | Plot |
|---|---|---|---|
| Estimated number of people who did *low or no exercise* | M1: $\mathrm{Y_i}$=66.65 + 1.47$\mathrm{X_i}$- 1.12$\mathrm{X_i^2}$ | 0.55 |  |
| Estimated number of *obese people* | M2: $\mathrm{Y_i}$=31.26-3.67$\mathrm{X_i}$ | 0.73 |  |

| | | | |
|---|---|---|---|
| Estimated number of people with *high blood pressure* | M3: $Y_i=22.56+0.63X_i$ | 0.04 |  |
| Estimated number of people with *psychological distress* | M4: $Y_i=12.82+ 0.38X_i- 0.34X_i^2$ | 0.27 |  |
| Estimated number of people who were *current smokers* | M5: $Y_i=17.98-1.85X_i$ | 0.53 |  |
| Estimated number of people who had *risky alcohol consumption* | M6: $Y_i=15,72-0.16X_i$ | 0.00 |  |
| Estimated number of people with a *risky waist measurement* | M7: $Y_i=65.07-2.80X_i$ | 0.67 |  |

Both M3 and M6 had tiny R^2 values, and the F-test performing on M3 gave a p-value of 0.189. We can then easily conclude that the density of facilities has no significant impact on people's high blood pressure and risky alcohol consumptions.

From M2, M5, M7, facility density had explained more than half the variance of the independent variable. Density of facilities in each area is statistically significant in fitting the proportion of obese people, smokers, and people with a risky waist measurement in that area.

Moreover, both M1 and M4 generated p-value less than 10^(-4) in F-test. Both the number of people doing low exercise and the presence of psychological distress have a fairly strong quadratic relationship with the building of facilities. According to the plot, the number of people suffering these two risks presented a rapid decline trend with the increase in facility density.

Regression results imply that increase in the number of facilities per km^2 may contribute to improvement in these five health risk factors (except alcohol consumption and blood pressure).

It provides a rationale for our hypothesis. But one issue is the reasonability of assuming a causation between these risk indicators and building of facilities.

## 3.3.Classification

Next, we use decision trees to do classification to see whether we can classify the level of people's health risks based on the amount of sport facilities.

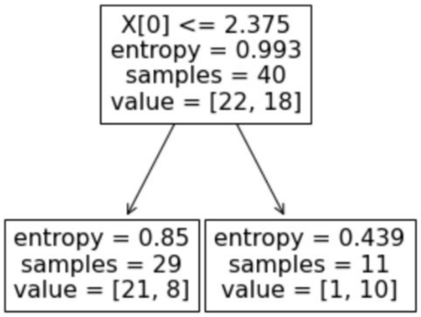
We use the sports facilities per square kilometer as the attribute for all models, and the health risk factors as class label for each model. To classify the level, we first find the group mean value of each amount of disease. Next, for those records equal or above the mean we mark them as 'high'; else, mark them as 'low'. Then, we run the test. Therefore, the model will be like

$$X_i = \textit{the sports facilities per capita per square kilometer} \text{ for } i = 1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7$$

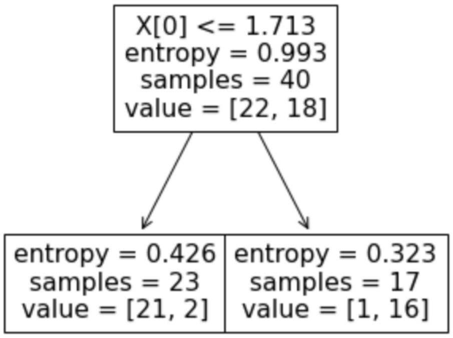
And

$$Y_1 = \textit{the level of estimated number of people who did low or no exercise}$$
$$Y_2 = \textit{the level of estimated number of obese people}$$
$$Y_3 = \textit{the level of estimated number of people with high blood pressure}$$
$$Y_4 = \textit{the level of estimated number of people with psychological distress}$$
$$Y_5 = \textit{the level of estimated number of people who were current smokers}$$
$$Y_6 = \textit{the level of estimated number of people who had risky alcohol consumption}$$
$$Y_7 = \textit{the level of estimated number of people with a high risk waist measurement}$$

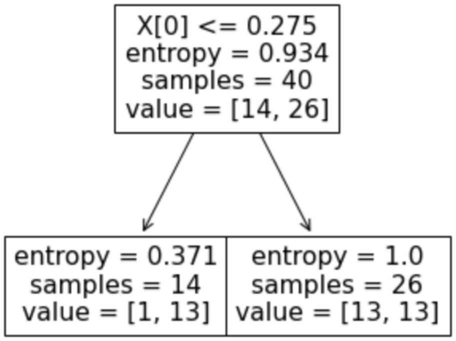
M1:


X[0] <= 2.375
entropy = 0.993
samples = 40
value = [22, 18]
entropy = 0.85
samples = 29
value = [21, 8]
entropy = 0.439
samples = 11
value = [1, 10]

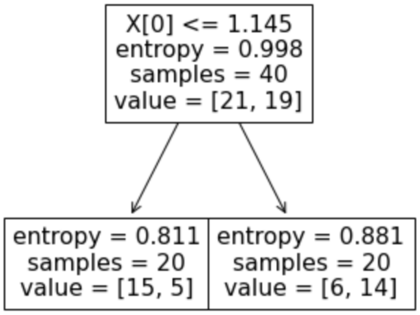
M2:


X[0] <= 1.713
entropy = 0.993
samples = 40
value = [22, 18]
entropy = 0.426
samples = 23
value = [21, 2]
entropy = 0.323
samples = 17
value = [1, 16]

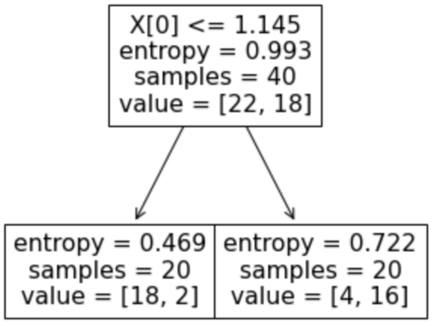
M3:


X[0] <= 0.275
entropy = 0.934
samples = 40
value = [14, 26]
entropy = 0.371
samples = 14
value = [1, 13]
entropy = 1.0
samples = 26
value = [13, 13]

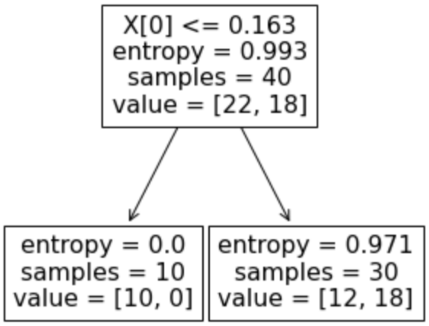
M4:


X[0] <= 1.145
entropy = 0.998
samples = 40
value = [21, 19]
entropy = 0.811
samples = 20
value = [15, 5]
entropy = 0.881
samples = 20
value = [6, 14]

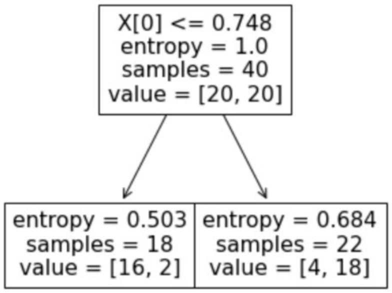
M5:



M6:



M7:



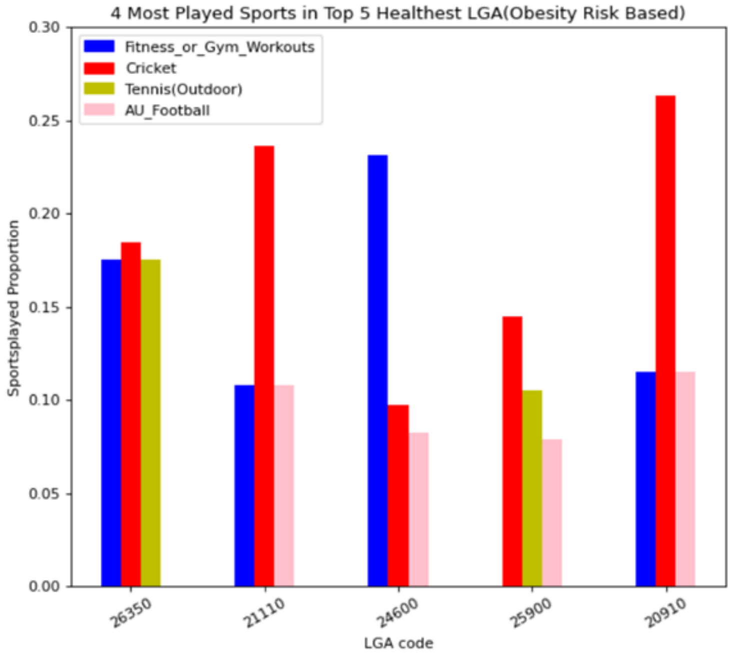
Using entropy as criteria, we have found out that, except for model 3 and model 6, rest of the models have relative low entropies in the classification indicating that based on the level of amount of facilities, we can produce reasonably accurate prediction of the level of health risk.

## 3.4.Group by Type

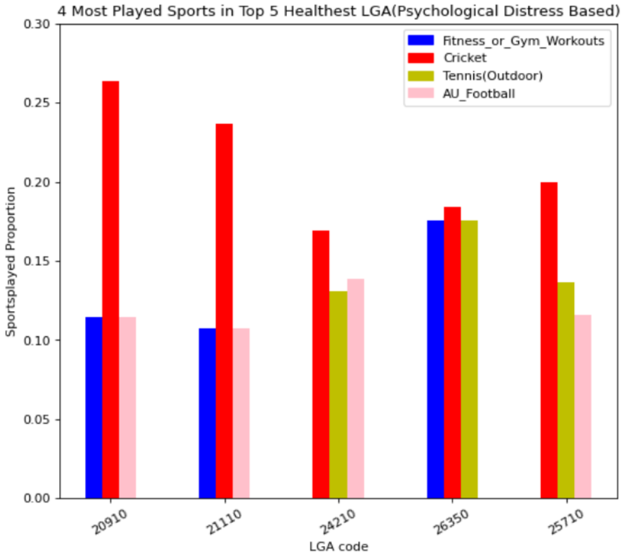
To further supplement our analysis, for each risk measure that proved to be correlated to facilities, we have selected five LGAs with the best performance and listed the three sport types with the highest proportion in each LGA.

Here for simplification, we combined three obese-related indicators 'obese_p_2_asr', 'lw_excse_2_asr', 'wst_meas_p_2_asr' into one called 'obese_risk'.

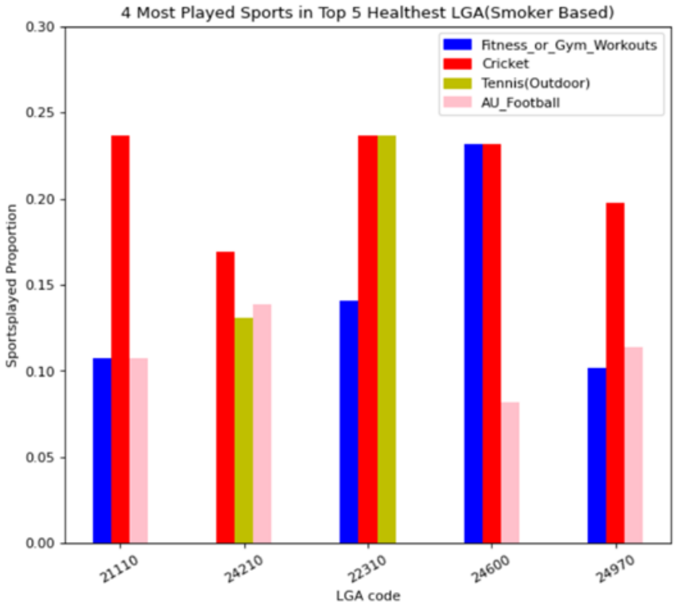
Obesity Risk:



Psychological distress:



Smoking:



From the bar charts, all LGA with best health profile are dominated by fitness or gym workouts, cricket, tennis, and AU football. And the cricket seems the most popular one in those healthy areas. This provides a hint for facilities development in those poor wellbeing areas.

## 4. Discussion

The major limitation is that correlation can suggest an underlying causal relationship but does not necessarily imply it. Even though a strong and reliable correlation was observed from our analysis, there is no guarantee that development of sport facilities can effectively improve people's willingness to exercise and mitigate people's health risk. Maybe other factors such as income level or education are defining people's lifestyle, and the number of sports facilities are simply a reflection of that. This leads to the following paragraph about the future improvements for this project.

When looking at the project retrospectively, we realized that the distribution of age, income and education could be distinct among different areas. Since we do not take the demography into considerations, the potential correlation with health condition is ignored. This can be improved by including those variables of each LGA in a multiple regression and using relevant tests to check which variable is significant in the model.

## 5. Conclusion

As our simple linear regression model and decision tree suggests, there is almost certainly correlation between the density of sport facilities and percentage of people with health risk factors. Hence for local government areas with a low density of sport facilities, it might not be a bad idea to make more parks and reserves available to the public which will incentivize more people to exercise, and ultimately improve the overall healthiness and livability of the community.