**COMP20008 Assignment 2**

**Project Proposal**

**1.**

Does the number of adults per sporting facility in a LGA (Local Government Area) affect the rates of obesity per 100 adults and the number of overweight adults per 100? This research question relates to both the themes of liveability based on the amount of available sporting facilities there are within the suburb and the theme of health to examine if the number of adults per sporting facility correlates with the rates of overweight and obese people within that LGA.

**2.**

Local councils may want to use this data to discuss if they need to invest in more local sporting facilities if it is found that there is a correlation between the weight of their adult citizens and the number of adults per sporting facility. People potentially moving into a new area may also find the data useful to see if they may be at risk of gaining weight due to a higher rate of adults per sporting facility in the LGA they are looking at moving too.

If the results came back that having more than a certain rate of people per facility in a LGA causes higher rates of obesity and higher rates of adults being overweight a statewide minimum rate could be implemented to ensure the overall health of the state and reduce statewide numbers of obesity.

**3.**

In order to answer the hypothesis question, three datasets have been identified. The first dataset was collected from AURIN and is titled Victoria Sports and Recreation Facility Locations for 2015 & 2016. This is quite a large dataset consisting of 9471 rows and 19 columns. The data comes in csv as well as json format. The large size of the dataset is due to the fact that the information is available in the granularity of suburbs. Most of the 19 attributes are not required for our task. The ones that are important to us include Local Government Area (LGA), facility type, and the kind of sports played.

The second dataset was also gathered from AURIN. It contains information about Adult Health Risk Factor Estimates. Keeping in line with AURIN datasets, it comes in both csv and json format. This dataset is substantially smaller than the previous one containing only 80 rows and 8 columns, the attributes provided are LGA, weight measurement, exercise rate, and obesity rate. The third dataset is the adult population per LGA with 80 rows. The datasets can be combined through the common attribute - LGA.

**4.**

Any part of the dataset that is with text needs to be pre-processed. This begins with similar columns that need to be compared such as LGA name which is on both datasets. To ensure each record is matched to a similar record from the other datasets, the pre-processing that occurs will consist of converting all to lowercase, removing any types of spaces or punctuation like hyphens and then to compare them to the other dataset to see if they are the same thing.

For analysis on the health conditions of individuals, a tally of which sports played and adults within the LGA and then this information can be compared against the health information of individuals to provide estimates.

**5.**

The initial preprocessing can provide a link between the 3 datasets so that the information in the three datasets can be related to one another. The analysis information can help determine how each sport and number of adults within an LGA may affect certain health conditions. Graphs can be generated based on this information to represent obesity against adults in an LGA and another to represent overweight individuals against adults per sporting facility.

**6.**

One of the challenges the team will face during this project is having 3 people do the work of 4 as one of the members dropped out. As for the work itself, the first obstacle would be to clean and organize the data. Since the datasets do not have any time attributes, an assumption has to be made that the information remained the same from 2014 to 2016. There is a risk that insights found in the project are due to factors that have not been considered or taken into account.