Data Science Project – Group 146
Diabetes & Recreational Facilities

What is the relationship between the prevalence of diabetes and recreational facilities in Victoria?

1. The Research Question

This question relates to the health of communities in Victoria. Diabetes is the most common health related disease, and recreational facilities such as: gyms, parks, sports grounds et cetera are expected to benefit the health of individuals in a community. With this data we can better understand the health of these communities and whether there is a correlation between the two variables.

2. Why This Question

Understanding the relation between diabetes and recreational facilities can help reduce the risk of individuals in certain areas from developing type 2 diabetes. This kind of information would be useful for health officials as well as local councils because it could help councils to decide what kind of facilities should be built in areas where cases are particularly high to improve the general health of the population in that area.

It might provide innovative information because if it is found that areas with more recreational facilities tend to have lower cases of diabetes, some information we can deduce is that the population will tend to use recreational facilities if they are available and that the issue in some areas may be that there are not enough recreational facilities made available.

3. Data Sources

The two open datasets below can be linked together as they both include a location field, and thus we can relate the accessibility of recreational facilities and the prevalence of diabetes in different locations within the state of Victoria.

Data Vic: Sport and Recreational Facilities List

The Sport and Recreational Facilities List is a Victorian government dataset which identifies all recreational facilities within the state in terms of the type of sport played, facility purpose, condition, age and location, in addition to other details. The dataset is formatted as a Microsoft Excel file (XSLX) and is 3.7 MB in size.

NDSS: Australian Diabetes Map

The Australian Diabetes Map is a national map monitoring the prevalence of diabetes in Australia, and shows people diagnosed with diabetes who are registered on the National Diabetes Services Scheme. The Australian Diabetes Map reflects diabetes prevalence in Australia in terms of location at the national, federal, and state electorate, local government area, and postcode levels. Additionally, the map allows you to download data for these locations in either the CSV or JSON formats.

4. Methodology

This study will utilise various data wrangling techniques to clean and massage the data into a suitable form for the analysis. The analysis will report the correlation between diabetes population and access to recreational facilities across Victoria.

The team will wrangle the data from *Australia Diabetes* to target the intended population where the scope of this study is resided in the state of Victoria. The Victoria diabetes population data will be processed further to provide the granularity on the suburb and postcode level.

The team will collect the recreational facility data from the Department of Health, Sport and Recreation Victoria, and Victoria Government databases. The data across different agencies will be filtered and processed before they are added into an aggregated dataset. This dataset will provide an understanding on the level of accessibility on the aspect of recreational facility.

The analysis will visualise statistical data and map variables geographically to present the pattern between diabetes and recreational facility. The statistical summary will be provided to give further details on the visualised patterns. The methodologies – visualisation and statistical analysis – will help the team to conduct investigations into the subject matter.

5. What can be achieved by this Methodology

By using these data wrangling methods, we can find a correlation between the access to recreational facilities in an area, and the diabetes population proportion in that area, where area is defined to be at the postcode level. We propose the output of the data wrangling to be pre-processed data, in the form of a scatter plot, which maps the access to recreation centres, to the number of diabetes cases. We would also continue to present statistics on different diabetes types, since not all are related to the same causes, such as genetics or other external factors.

This adds value because the raw data alone, is not sufficient to deduce any conclusion, if we combine the raw data from various sources, and pre-process the data, we can use the data to find some conclusions.

6. Challenges and Limitations

Some limitations of this work are that although there may be a visible relationship between diabetes and recreational facilities, there may be other factors leading to higher diabetes rates, such as certain areas having more fast-food stores, and so our conclusion may provide incorrect information if the data is not determined to be directly correlated.