COMP20008_2021_SM1
Assignment 2: Data Science Project

Group 146:

1068125	Ed Yi-Hsuan Peng
1181481	Jett Miller
1187107	Brandon Widjaja
1143743	Zimo Peng

The impact of the accessibility of sport and recreational facilities on the prevalence of diabetes in Victoria's local government areas

Contents

Introduction

Diabetes is a prevalent disease worldwide which affects 1 in 11 adults (Diabetes Australia, 2021). According to Australian Institute of Health and Welfare, 4.9% of people are affected by diabetes nationally (2018). Diabetes remains a major interest in health studies due to people with diabetes being vulnerable to various health risks (Roglic, 2016). Health studies found that individual lifestyle practices play a part in developing type 2 diabetes, thus altering maladaptive lifestyles can prevent and mitigate the disease (Harvard Health Publishing, 2021). It is clear from the available evidence that diabetes can be managed by maintaining regular exercise (Sato et al., 2007). Therefore, this study will focus on type 2 diabetes because it composes 95% of total diabetes cases and it is believed to be largely caused by environmental factors (Bi et al., 2012).

This data analysis report investigates the relationship between diabetes and physical exercise in relation to the accessibility of sport and recreational facilities. The study is focused on how the sport and recreational facility unit count can impact people's physical activity habits and sports participation rates in each Victorian local government area (LGA), hence influencing the diabetes rate in each Victorian community. This report will give insights into the interplay between public health and infrastructure, specifically in the subject matter of the diabetes disease and facility accessibility to understand Victorian's health and Victorian communities' liveability and inclusiveness. This analysis report will be valuable to inform public policy and health intervention strategies.

Materials

The data used in this data analysis are the following:

- *NDSS Diabetes Map*
    - National Diabetes Services Scheme (NDSS), Australia
    - https://map.ndss.com.au/
- *Sport and Recreational Facilities List*
    - Data Vic, Victoria's open data platform
    - https://discover.data.vic.gov.au/dataset/sport-and-recreational-facilities-list
- *2015 Local Government Area Profiles*
    - Department of Health and Human Services (DHHS), Victoria, Australia
    - https://discover.data.vic.gov.au/dataset/2015-local-government-area-profiles

These three datasets are complete and consistent in fields which are relevant to our study. The datasets are highly believable and easily interpretable; however, the timeliness of datasets may be improved upon. The diabetes data, which we downloaded from each LGA's diabetes profiles in April 2021, was pre-processed out from *NDSS Diabetes Map* using "process-diabetes.py". The facility data will be pre-processed out from the *Sport and Recreational Facilities List,* a dataset which was last updated in August 2020, using "process-facilities.py". The lifestyle and health related data will be pre-processed out from *2015 Local Government Area Profiles*, a dataset last updated in March 2018, using "process-govt_profiles.py". The data quality –timeliness– is a constraint in our analysis, however, we

ensured to use the most recent diabetes dataset available at hand and strived to minimise the time difference between inter-datasets.

We pre-processed each original dataset using pandas, re and os libraries, then outputted each data frame to a comma-separated values file format (CSV) with the key being Victorian LGAs. After having three dataframes in CSVs, we linked the data by joining on their shared key values –LGA– into a collective dataset with various fields of interest using the script: "LGA_diabetes_profiles.py". The dataset obtained is called "LGA_diabetes_profiles.csv"; this dataset will be used for further investigations and analyses.

Methods

We used matplotlib to visualise our data by creating multiple scatter plots using the prevalence of diabetes as the dependent variable and various independent variables. Furthermore, we classed each datapoint as belonging to a certain type of LGA, including a shire, rural city, city, or borough, which helped to distinguish trends for rural and urban areas regarding type 2 diabetes.

We decided on using a simple linear regression to analyse the data that we collected to estimate the relationship between the variables, as we expected some linear relationship between the prevalence of diabetes and the variables. We chose to use the simple linear regression over an alternative such as multiple linear regression as we were examining the relationship between the dependent variable and one independent variable, rather than altogether. Furthermore, the independent variables we were using were likely to depend on one another; such dependencies violate an assumption of multiple linear regression.

We performed the residual analysis on each plot to confirm that using the simple linear regression to model the relationships was appropriate. The residual analysis was essential to determine whether the model's prediction was suitable by checking that the errors were random and independent.

While we recognise that outliers can have a negative effect on regression analysis, we were interested in the granularity of each local government area and each data point was important to help discern trends for the different LGA classes.

Results

Note. The LGA status: 0, 1, 2, 3 are corresponding to shire, rural city, city, borough respectively.

Figure 1

Diabetes rate in percentage to four influencing factors.

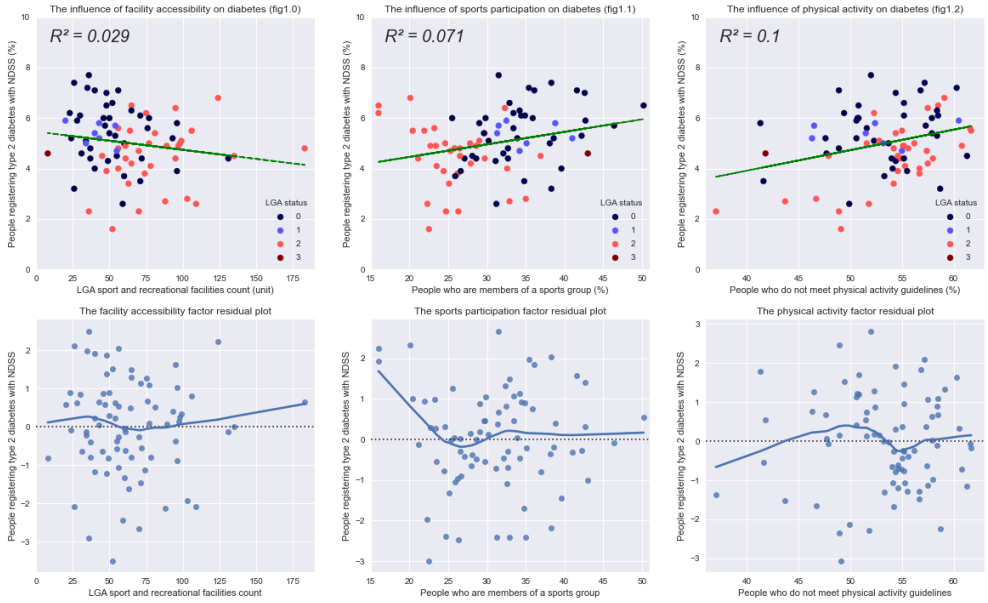

Figure1.0 demonstrates a weak negative correlation. The Pearson correlation coefficient (R) is 0.17. The residual plot is smooth for fig1.0. The linear regression line is $\boldsymbol{y = 0.0072x + 5.4647}$. **Figure1.1** demonstrates a moderate positive correlation. The R(fig1.1) value is 0.266. The residual plot is influenced by outliers for fig1.1. There are two class-groupings to be observed. **Figure1.2** demonstrates a moderate positive correlation. The R(fig1.2) value is 0.316. The residual plot is influenced by both outliers and a small cluster on the x-axis around 55.

Figure 2

The influence of facility unit count on physical and sporting activity.

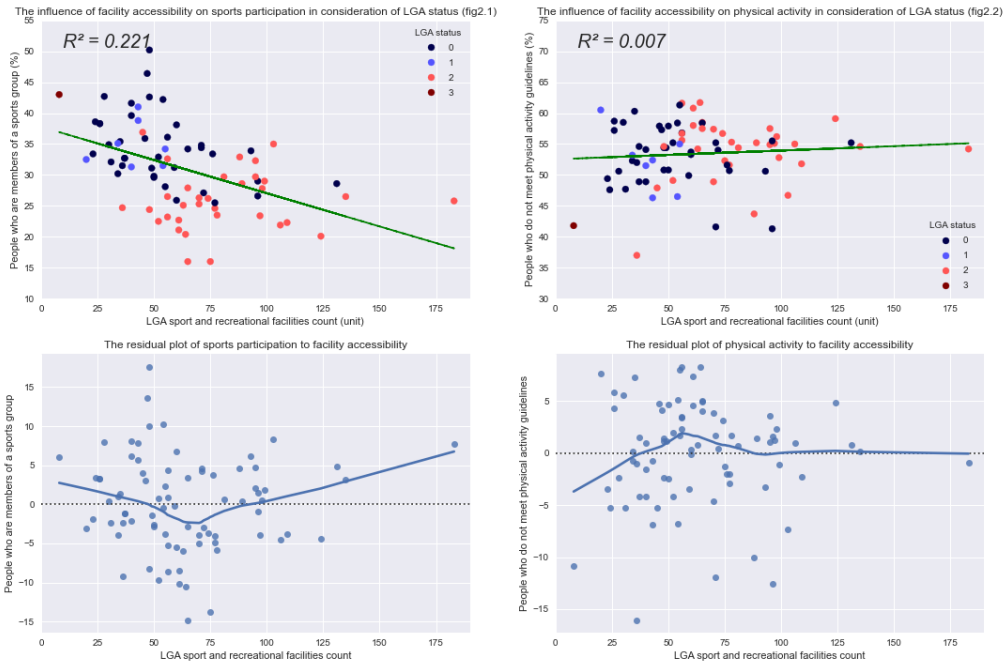

Figure 2.1 demonstrates a strong negative correlation. The R(fig2.1) value is 0.47. The residual plot is influenced by an outlier. There are two class-groupings to be observed. **Figure2.2** demonstrates a weak positive correlation. The R(fig2.2) value is 0.084. The residual plot is influenced by outliers.

Discussion

The influence of facility accessibility on diabetes (fig.1.0) has a very weak negative correlation between the two variables. Through the residual analysis, it shows that the model fits well as the observed and predicted values differ only by a small margin. The model presents a considerably low R^2 value, so it is unlikely that the facility count has a substantial impact on the amount of diabetes registrants in an area. The slight correlation represented by the R value is a valuable indication of some relationship between the two variables, although there is no well-established direct impact between the two. Therefore, we conducted a further investigation into the matter and produced **Figure2** for extended analysis.

The influence of sports participation on diabetes (fig.1.1) depicts a weak to moderate positive correlation between the two variables. The residual plot is not smooth due to being affected by two to three outliers at the lower end of the x-axis, otherwise the model fits well where points are randomly dispersed. There was significant class grouping (fig.1.1) by the LGA status which illustrates that people in cities were less likely to participate in a sports group than those in shires. It is important to note that type 2 diabetes is more prevalent in shire communities as observed from **fig.1.1** and **fig.1.0** as blue-coloured dots (status=0,1) tend to position higher on the y-axis as a class-cluster group.

The influence of facility accessibility on sports participation in consideration of LGA status (fig.2.1) is moderate negative correlated as shown by the best-fit line. The R^2 value illustrates a decent suitability of the linear regression to the data points; furthermore, the residual analysis demonstrates that the linear regression is considerably appropriate with some outlier influence. Here, we observed a considerable class-grouping between blue-coloured (status=0,1) and red-coloured (status=2) LGAs. The class grouping suggests that shire and rural communities have a culture of participating in sport-clubs, whereas cities are less involved in sport-groups as observed from the graph. Moreover, cities have more facilities due to the larger demand/population, whereas shires and rural cities may have less incentive to open-up facilities. The consideration of facility type is important in interpreting this result because different facilities may encourage group-sports participation more than others. Thus, it needs a further study to understand LGAs' facility types.

The influence of physical activity on diabetes (fig.1.2) has a positive correlation. The residual analysis shows that the data is symmetrically distributed, except where a small cluster groups around 55 on the x-axis and the outlier at the lower end. Thus, the result obtained here suggests that physical activity may reduce diabetes.

The influence of facility accessibility on physical activity in consideration of LGA status (fig.2.2) has a trivial relationship between the variables, which indicates that they may be independent. This suggests that people conduct "physical activity" regardless of the number of facilities around them. This may be explained by the diminishing returns of the infrastructure where more facilities do not bring more people to exercise. Furthermore, the issue may also lie in the definition of "physical activity" set by DHHS, where the guideline requirement can be achieved in numerous ways without needing to access any facility.

These results are valuable because the link between facility count and the prevalence of type 2 diabetes is not immediately evident. However, in conjunction with our supplementary results, we deduced several interpretations. Our results show that LGAs with a higher

proportion of people who meet physical activity guidelines correspond to a lower proportion of diabetes registrants. Furthermore, the results illustrate that the proportion of people meeting physical activity guidelines is independent of the number of facilities. This finding, which is evidenced by our model, demonstrates that the increase in the number of facilities is unlikely to lower the number of type 2 diabetes registrants in an LGA.

Limitations & Further Studies

One limitation of the study was that the data collected had time differences, therefore the analysis could not produce the most up-to-date prediction model. The data timeliness could be improved upon to give the most relevant advice to policy makers and health officials.

Additionally, this study did not specify the type of facility in the count. Our results illustrated significant class-groupings in the LGA status; therefore, it may be possible that the type of facility in each LGA is disproportional in some regards, which could have skewed the results.

As a result of the class-grouping phenomena which is reported in the analysis, future studies should investigate the subject matter in terms of each LGA's demographic profile. We hypothesise that the class-grouping is caused by demographic attributes such as: age or socio-economic status. It is important to learn the underlying reason for the class-grouping, hence the insights can be offered to LGAs to tackle diabetes.

Conclusion

The report concludes little evidence that the increase of facility count will lower the diabetes rate in Victorian local government areas; however, the physical activity illustrates some influence in that regard. Upon a further examination, the analysis establishes that facilities do not influence people's physical activity habits and the prevalence of diabetes is tied to each local government area status.

References

Bi, Y., Wang, T., Xu, M., Li, M., Lu, J., Zhu, X., Ning, G. (2012, December 28). Advanced research on risk factors of type 2 diabetes. *Wiley. 28*(2). 32-39. https://doi.org/10.1002/dmrr.2352.

Data Vic. (2018, March 09). *2015 Local Government Area Profiles* [Data set]. State Government of Victoria. https://discover.data.vic.gov.au/dataset/2015-local-government-area-profiles

Data Vic. (2020, August 28). *Sport and Recreational Facilities list* [Data set]. State Government of Victoria. https://discover.data.vic.gov.au/dataset/sport-and-recreational-facilities-list

Diabetes Australia. (2020, December 31). *Australian Diabetes Map* [Data set]. National Diabetes Services Scheme. https://map.ndss.com.au/

Diabetes Australia. (2021). *Diabetes Globally. https://www.diabetesaustralia.com.au/about-diabetes/diabetes-globally/*

Harvard Health Publishing. (2021, February 03). *The importance of exercise when you have diabetes. https://www.health.harvard.edu/staying-healthy/the-importance-of-exercise-when-you-have-diabetes*

Roglic, G. (2016, June 29). WHO Global report on diabetes: A summary. International Journal of Noncommunicable Diseases. 1(1). 03-08. https://www.ijncd.org/article.asp?issn=2468-8827;year=2016;volume=1;issue=1;spage=3;epage=8;aulast=Roglic

Sato, Y., Nagasaki, M., Kubota, M., Uno, T., Nakai, N. (2007, January 29). Clinical aspects of physical exercise for diabetes/metabolic syndrome. Elsevier. 77(3). 87-91. https://doi.org/10.1016/j.diabres.2007.01.039