Assignment 2 Final Report for Elements of Data Processing

Group 118

Yufeng Xie 1166106   Hongfei Wang 1166435

Haozhe Liao 1166399   Yage Guo 1166008

1. Question & Theme

The research question of our project is which suburbs in Melbourne need new sports facilities. And this project is really good for the health level of people in Melbourne. More sports facilities always can affect the amount of physical activity a person participating in. As we all know, a person's health level is closely related to his frequency of exercise. If we can figure out which suburbs in Melbourne lack sports facilities, we can provide better medical services and build more sports facilities to these suburbs, which certainly can improve the health of the people in these areas. In addition, participation in sport is a human right. Everyone has the right to practice sport regardless of race and gender. Our project contributes to enable everybody to have the opportunities to access sporting facilities. This reflects the theme of inclusiveness in our project. As a result, this project is feasible and advantageous, which is really beneficial to the health and inclusiveness of people in Melbourne.

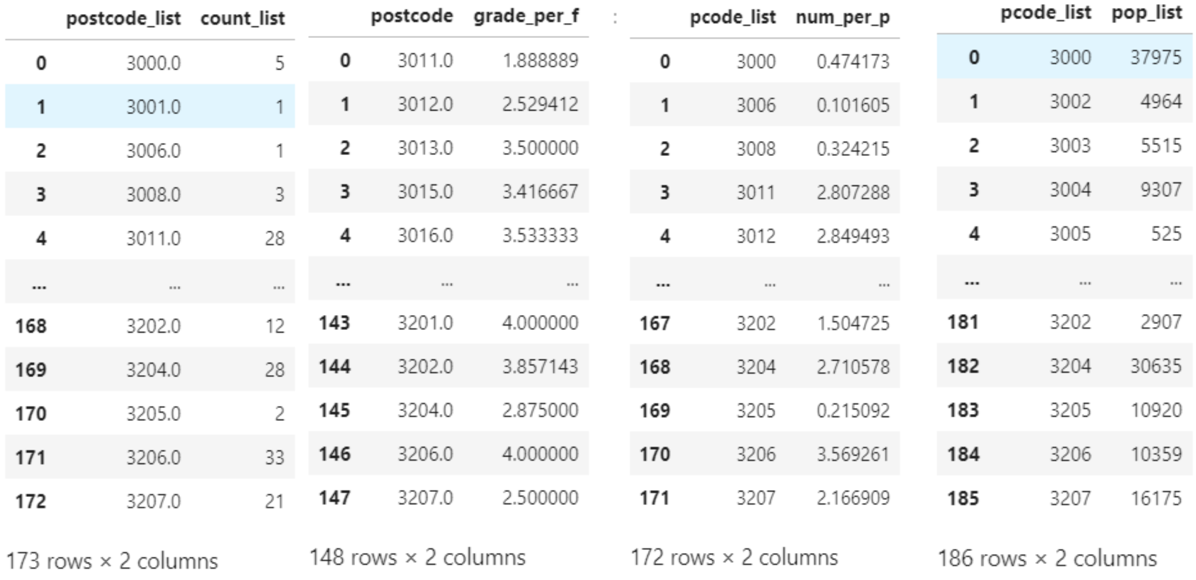
2. Datasets

We use three datasets in total. The first one is from AURIN DATA. It shows the number of the sports facilities in different suburbs in Victoria with postcodes provided. Also, there is data about the facilities' sports type, field type and some other features. The second one is also from AURIN DATA. It describes the facility condition and facility age with respect to different postcodes. The last one comes from Australian Bureau of Statistics, which contains data about the population in Victoria in 2016. The reason why we choose these three datasets is that the number of the sports facilities and its condition implies whether or not the region lacks sports facilities or needs to be replaced with new facilities. Choosing population data is to see if the number and condition will be influenced by the population factor. Since the postcode can simply refer to different suburbs, we will define postcodes as our fixed research objects and will pick key data such as the numbers of the facility, the condition of the facility and total population as our variables to carry out the following research and analysis.

3. Wrangling and analysis methods

All of the datasets we found contain data about the whole Victoria. In order to get the data of Melbourne, we first sort the data frames with postcodes starting from 3000 to 3207. There are two main aspects we consider about the question ‘Which suburbs need NEW sports facilities? ’: those suburbs that lack sports facilities, and those that have very old or poor-condition facilities. For these two research directions, data are to be processed separately.

To learn which suburbs lack sports facilities, we aggregate the number of sports facilities in each suburb based on the first data frame (see data frames in Graph 1). This gives us a list that shows the distributions of the total number of facilities of different postcodes in Melbourne, by which we form a bar chart (Graph 2). This graph indicates that some suburbs have more sports facilities than others. (Detailed result analysis is presented in the next part.) However, it is not rigorous to answer our question by simply considering the total amount. Hence next, we preprocess the data from the dataset of population in Melbourne and make a scatter plot illustrating the population of each suburb (Graph 3). From this plot, a conclusion can be made that Melbourne’s population distribution is generally balanced. Therefore, we decide to further calculate the number of facilities per person in each postcode in order to get the most comprehensive and reliable result. While the population is too large to be the denominator, we use math.log () function to preprocess the population list. Then, for every postcode from 3000 to 3207, a new variable named ‘num_per_p’ is calculated. This set of numbers we get now represents the per capita index of sports facilities in each suburb in Melbourne. If a suburb with a large population has the same number of sports facilities as a suburb with a small population, the former one certainly needs new facilities compared to the latter one. Thus, the final data frame performs the relationship between number per person and postcode (Graph 1), which gives a better way to answer the question. The bar chart is produced as Graph 4.



	postcode_list	count_list
0	3000.0	5
1	3001.0	1
2	3006.0	1
3	3008.0	3
4	3011.0	28
...	...	...
168	3202.0	12
169	3204.0	28
170	3205.0	2
171	3206.0	33
172	3207.0	21

173 rows × 2 columns

	postcode	grade_per_f
0	3011.0	1.888889
1	3012.0	2.529412
2	3013.0	3.500000
3	3015.0	3.416667
4	3016.0	3.533333
...	...	...
143	3201.0	4.000000
144	3202.0	3.857143
145	3204.0	2.875000
146	3206.0	4.000000
147	3207.0	2.500000

148 rows × 2 columns

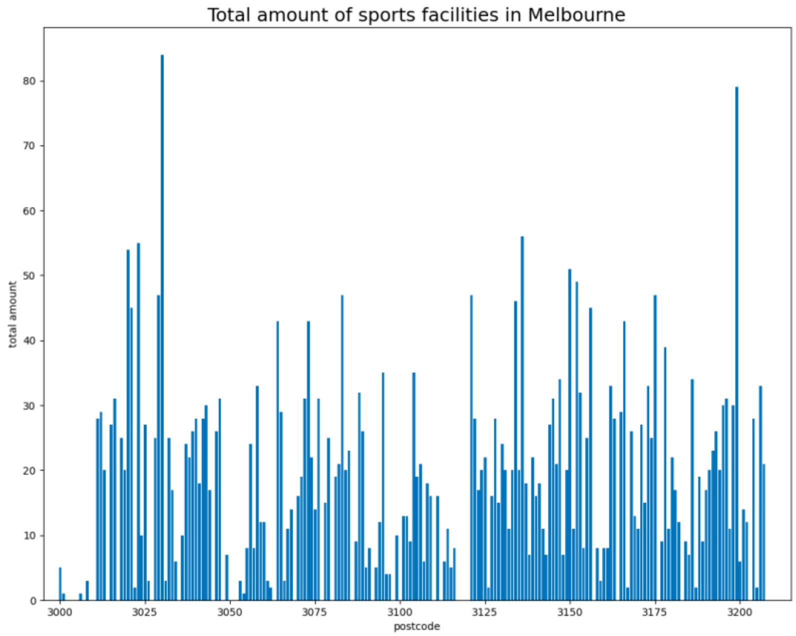
	pcode_list	num_per_p
0	3000	0.474173
1	3006	0.101605
2	3008	0.324215
3	3011	2.807288
4	3012	2.849493
...	...	...
167	3202	1.504725
168	3204	2.710578
169	3205	0.215092
170	3206	3.569261
171	3207	2.166909

172 rows × 2 columns

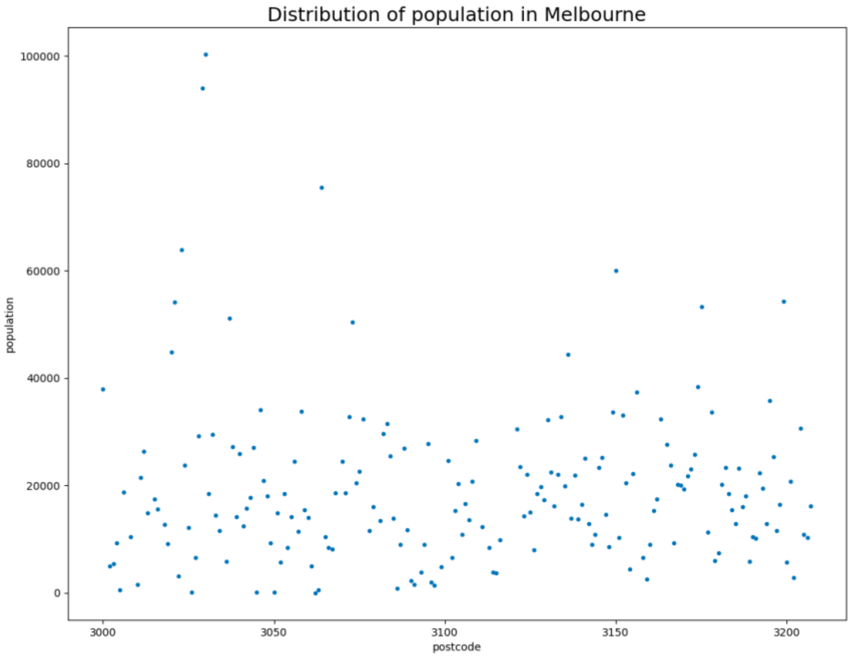
	pcode_list	pop_list
0	3000	37975
1	3002	4964
2	3003	5515
3	3004	9307
4	3005	525
...	...	...
181	3202	2907
182	3204	30635
183	3205	10920
184	3206	10359
185	3207	16175

186 rows × 2 columns

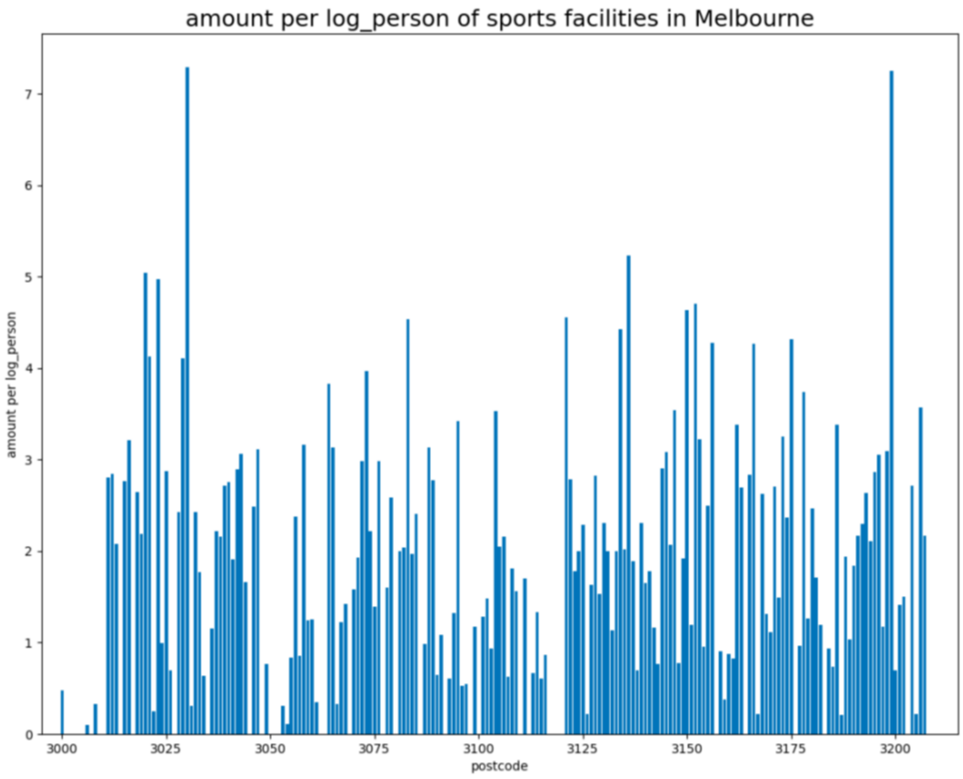
Graph 1: All data frames



Graph 2: Total amount of sports facilities of every postcode

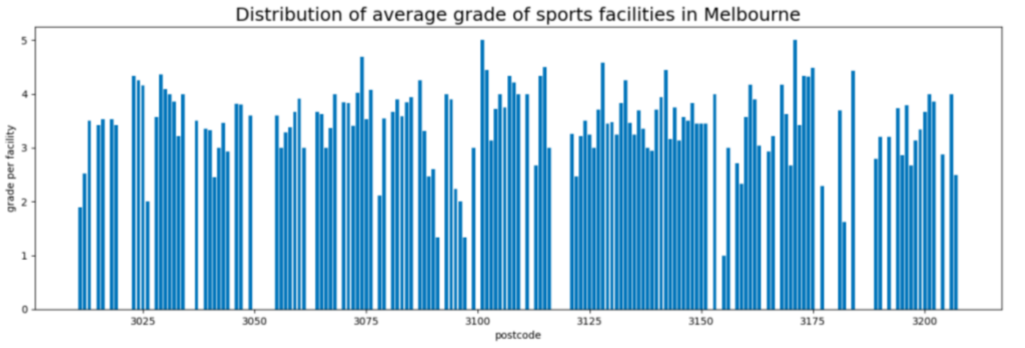


Graph 3: Population distribution of Melbourne suburbs



Graph 4: Number of facilities per log_person by postcode

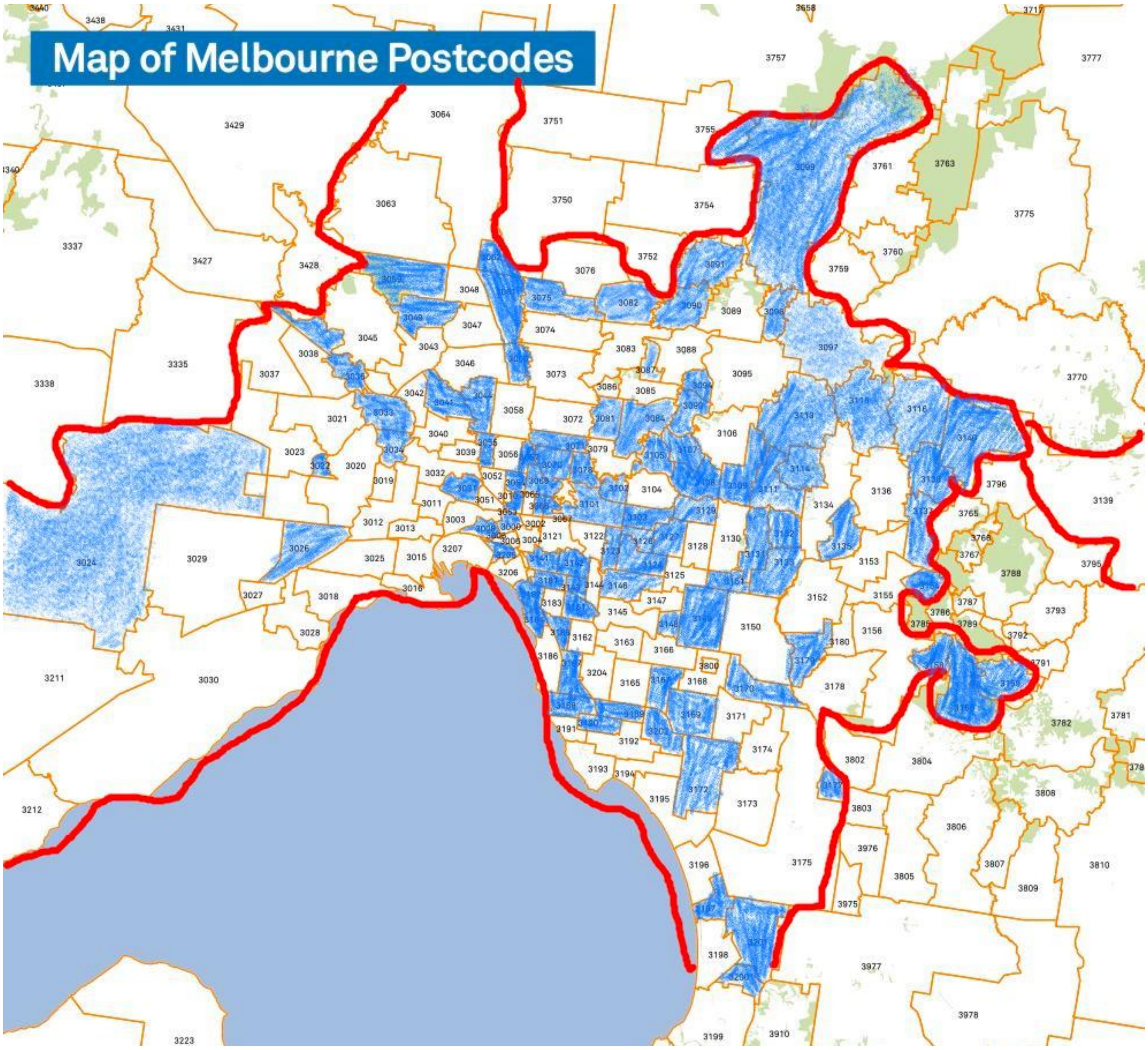
For the second aspect, we aim to find out the suburbs with many old or broken facilities. As the dataset 2 contains condition rates from 1(very poor) to 5(very good), we accumulate the total score of all facilities in each suburb, also sorted by postcode. Finally, we calculate the average condition score to produce a bar chart Graph 5.



Graph 5: Condition score of facilities by postcode

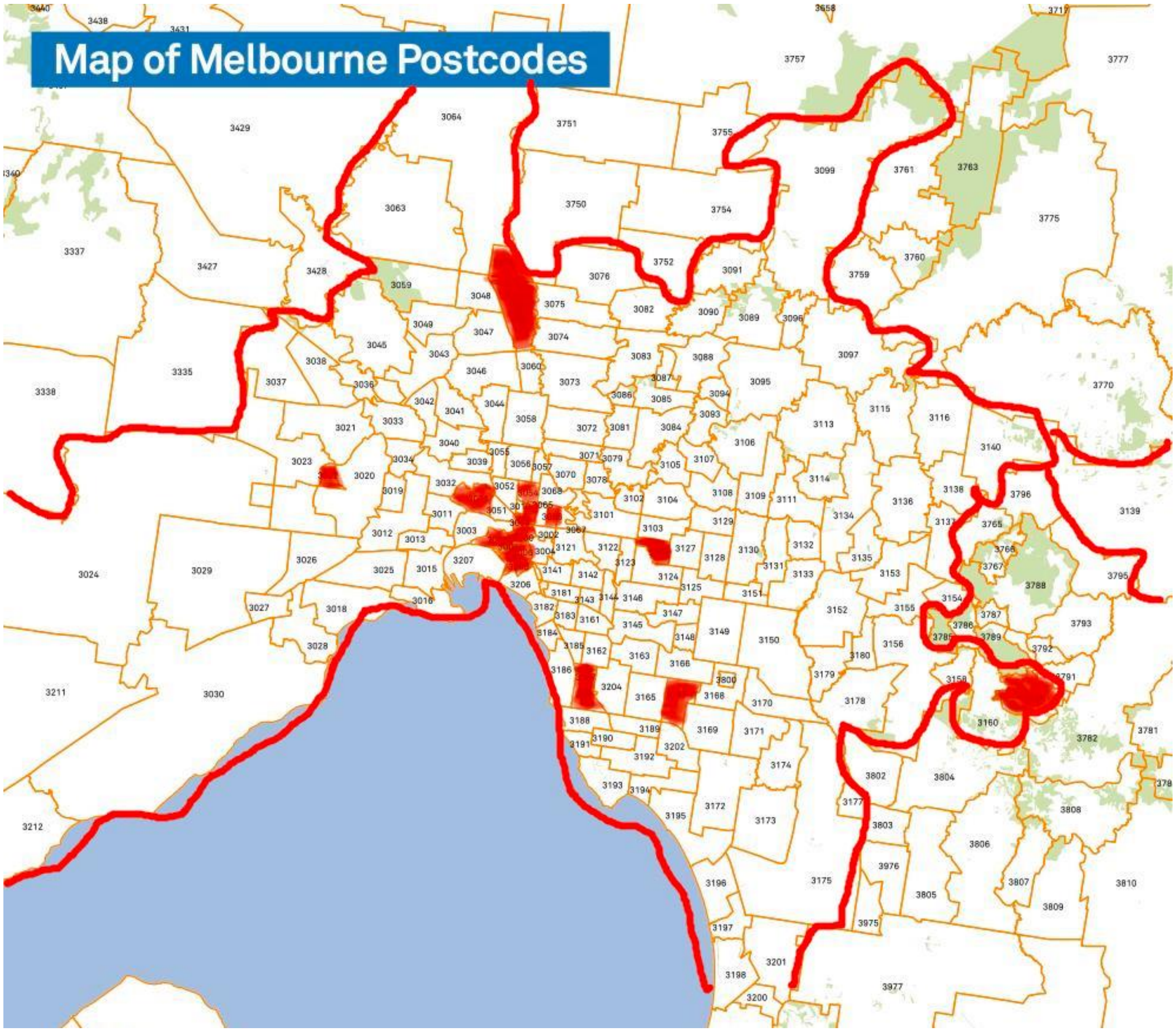
4. Key results obtained

From Graph (4), we see that the number of facilities per log_person varies between suburbs. To determine whether or not the region lacks sports facilities, we calculate the number of facilities that every log_person has, which is 2.07. So, we define this average value as our standard to determine whether or not the number of facilities is less than average level. To make it clearer where these regions are located in Melbourne, we do data processing again and pick out the postcodes whose value is less than 2.07 and then mark them in blue on a map with postcodes (see Graph 6).



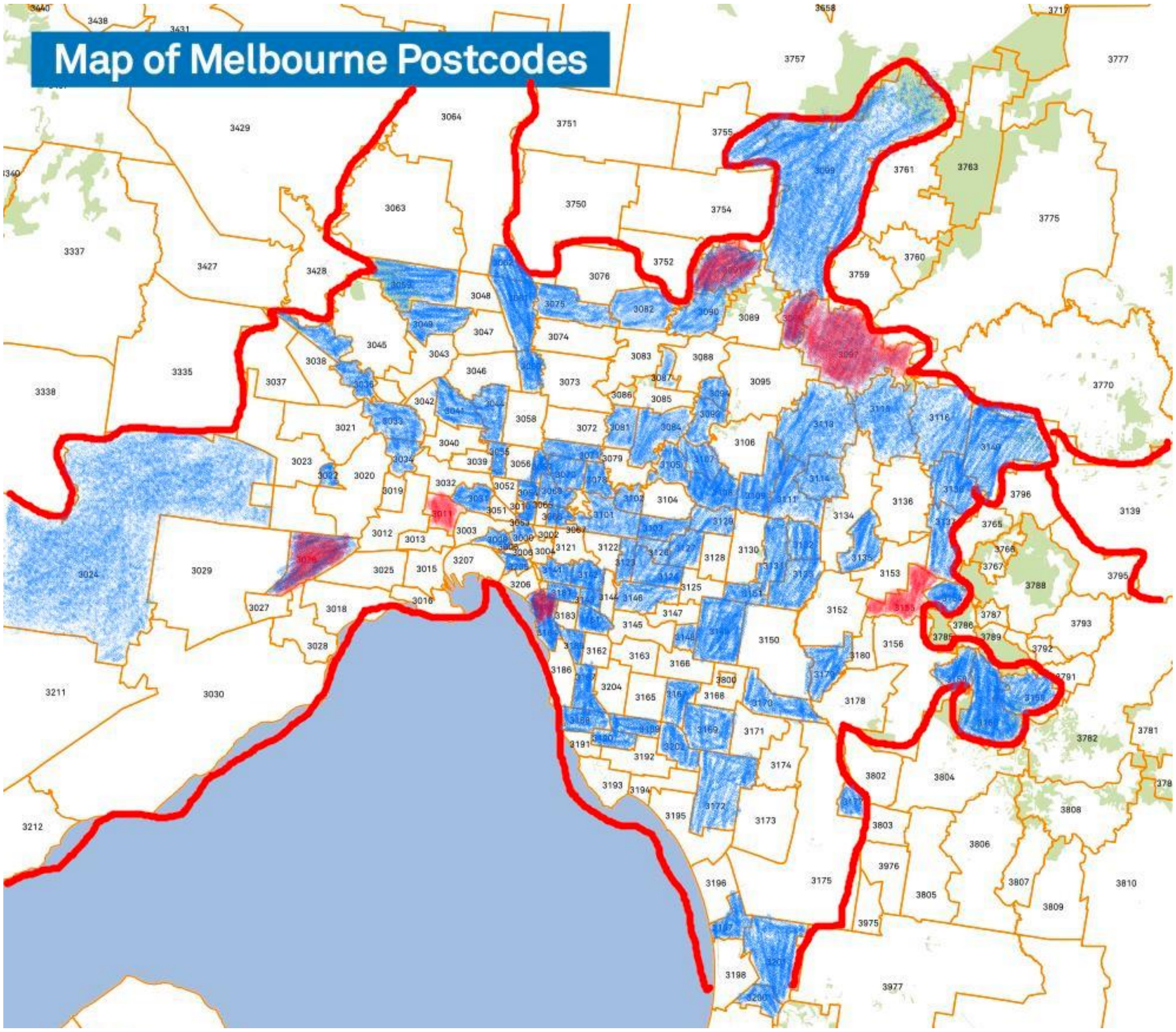
Graph 6: Suburbs with amount per log_person less than average

From the graph, we can generally look at the distribution of blue areas which are below the average level. However, since the blue areas cover over half of Melbourne (more than 90 suburbs), it is not apparent to determine which specified areas are short of sports facilities. To narrow the range, we choose the value 0.5 which means a great lack of sports facilities and pick out these regions and mark them red on the map (see Graph 7). Now we can clearly see that those areas in red are mostly concentrated around the centre of Melbourne city. Hence, we can draw a preliminary conclusion that the city of Melbourne needs more sports facilities. It is easy to explain why this happens, since most of the land in the city centre is used for business, working, education and housing, so it is quite difficult to divert more land for sports activities. This makes most people unable to enjoy the sports facilities and participate in sports activities.



Graph 7: Suburbs with amount per log_person less than 0.5

To combine the factors of the facility's condition, we first look at Graph 5. It shows that some of the suburbs have higher scores than others, so in order to make the result clearer, we use the same method of data processing and pick out 7 values from the data of the condition of the facility. These seven suburbs have the condition score of sporting facilities below 2 points, representing that the overall condition of sports facilities in this area is poor. We mark these 7 suburbs in red on the Graph 6 and get Graph 8. We can clearly find out there are 5 suburbs which are both marked by blue and red. These suburbs have not only poor-conditional sporting facilities, but also a low quantity of sporting facilities per person. Corresponding to the postcode on the map, these five postcodes (3026, 3091, 3096, 3097, 3182) represent five suburbs: Laverton North, Yarrambat, Wattle Glen, Bend of Islands and St Kilda. As a result, we consider that it urgent to input more new sports facilities into these five suburbs.



Graph 8: Suburbs with amount per log_person less than average (blue) and poor condition(red)

To conclude the result we have found, the centre of Melbourne severely lacks sports facilities and needs more new sports facilities to be constructed, and also for suburbs like Laverton North, Yarrambat, Wattle Glen, Bend of Islands and St Kilda, which are also short of new sports facilities and require more replacement for poor-condition facilities.

5. Significance of the results

The results we gain are significant and valuable for answering our research question. We spent quite a lot of time searching for appropriate datasets that may be helpful for our research. During the whole process of our project, we consider different aspects that might affect the results, and use various preprocessing and wrangling methods. For each part, various types of plots are chosen to best fit the data. To better present the results from our data, we annotated the map and led to the conclusion. Thus, the final results you see are reliable and valuable.

6. Limitations

First, only 2 available data sets about sports facilities were found. One is the zip code and the amount of equipment in the Melbourne area, and the other is the zip code in the Melbourne area and the service life and equipment conditions of the corresponding sports equipment. We encountered great difficulties when searching for data samples, which resulted in insufficient data samples found. This may be one of the limitations. Second, in the available data samples, there are a large number of missing or unrecorded values, which leads to a reduction in the scope of the research and brings a lot of trouble in the follow-up comparison and summary process. For instance, some of the postcodes did not have corresponding records. Third, it is hard to compare and exclude factors that have no data but can have an impact on the results of the study, only the impact of population distribution on the amount of equipment of different suburbs has been considered. Also, no relevant documents or records have been found as reference for the definition of the scope and standards of the lack of equipment.

7. Improvements in the future:

First of all, before proceeding to the next step of the research, try to find more data to support the argument, and the relevant factors that may have an impact on the research results should also find certain data for comparison and summary. Second, data processing should be more targeted. For example, when the per capita allocation of sports equipment is below what standard, new equipment should be considered, or new sports equipment needs to be updated when the level of new and old is below what standard.

8. Conclusion

In this project we researched the question of which suburbs in Melbourne need new sports facilities, corresponding to the themes of health and inclusiveness of people in Victoria. By preprocessing the datasets, we found and produced a series of plots and graphs, a conclusion can be made that the centre of Melbourne and other suburbs including Laverton North, Yarrambat, Wattle Glen, Bend of Islands and St Kilda, are short of new sports facilities and require more replacement for poor-condition facilities.