Phase 2B Submit Here

Print Options:

Include Questions & Answers Include Comments Include All Reviews Include File Info

Print

MAX PHILIP'S PEERMARK REVIEW OF ANONYMOUS'S PAPER

(0% COMPLETED)

"COMP20008 PHASE 2A" BY ANONYMOUS

COMP20008 PHASE 2A

'Determining a potential correlation between crime rates and sport facility availability'

Domain: Crime and sport

Motivation:

In this project, I will attempt to identify the degree to which there is a correlation between the availability of sport recreation centres, and the relative proportion of crime in that Local Government Area(LGA). Such information may be utilized by the State Government as an added perspective with regards to the factors that incite crime, and therefore may act as an additional consideration when creating a budget or introducing new policies.

Whether or not there is indeed a strong correlation, my project addresses the possibility that there may be a new way crime might be managed. The impact of sport, notably the ability for individuals to play it within their local area, as an alternative to illegal activity is not widely addressed and here I attempt to provide a more definitive answer with regards to this trend.

Datasets:

'Sport and Recreational Facilities' (Victorian Government Open Data Repository) - CSV: This dataset contains all sport facilities within Victoria. Furthermore, detailed about the location (notably the LGA) of the facility, as well as various other details such as sports played there, and number of fields are also noted.

URI: https://www.data.vic.gov.au/data/dataset/sport-and-recreational-facilities

'Crime Statistics Agency Data Tables – Crime by location' (Victorian Government Open Data Repository) – XLSX: The dataset records all crimes committed within Victoria between 2012- 2015. This too contains the LGA in which the crime was committed as well as various other details such as LGA population and the type of crime that was committed.

URI: https://www.data.vic.gov.au/data/dataset/crime-by-location-data-table

Purpose for 'wrangling':

Firstly, there is a lot of data that will not benefit me with regards to the comparison I wish to make, and therefore a large part of the wrangling process involves the removal of this data such that the only data that remains to be seen and analysed is that of relevance to my project. Such a process is vital here since the dataset I chose to use was not created specifically for my project and contains various other bits of detail that might only be handy when used for other purposes.

Secondly, the process of turning these datasets into graphs and charts further helps visualize if there are indeed any trends with regards to my variables. Here, even having data that is all completely relevant to the project may still be difficult to interpret if it is in the form of a huge table containing thousands of rows.

Integration:

The first thing I sought out to do was convert all datasets into the CSV format such that I would be able to easily import and wrangle the data within a Jupyter worksheet. This was already done for the 'Sport and Recreational Facilities' file. For the 'Crime Statistics Agency Data Tables – Crime by location' I simply created a new excel spreadsheet and copied the table over, saving it in the CSV format as required.

I also sought to make some minor changes to the spreadsheet itself within excel in order to make the process easier once I began within Jupyter. I renamed the columns, replacing spaces with '_' such that they may be more easily called upon, and replacing 'LGA ERP' to 'LGA_population'. Furthermore, there were a few '≤' under the column of 'offense_count' in the 'Crime Statistics Agency Data Tables – Crime by location' dataset that I chose to remove for ease of wrangling since it was always next to single digit numbers and the effect it would have on the results would be minimal.

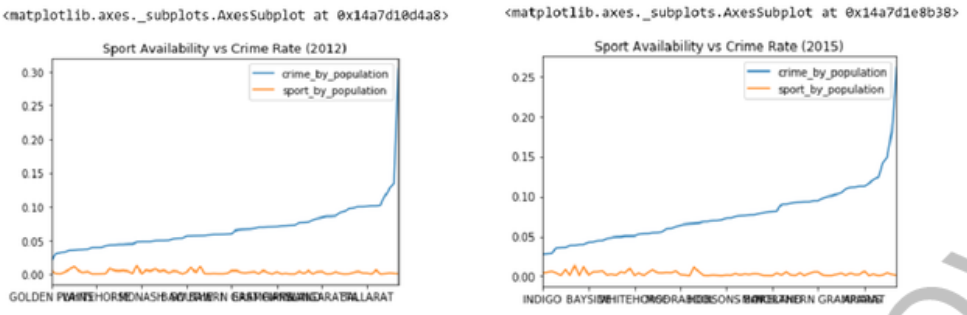
I then proceeded to create a new, modified table for the two datasets containing only the columns I thought were relevant. The modified table for 'Sport and Recreational Facilities' I called 'sport' and the modified table for 'Crime Statistics Agency Data Tables – Crime by location' I called 'crime'. The only column I retained for 'sport' was the LGA, since I was only interested in the number of sporting facilities in each given LGA. For 'crime' I retained the columns for 'year', 'LGA', 'offense_count' and 'LGA_population'. For now, I was only interested in the total number of crimes for any given area, regardless of how serious the offense might be.

I also decided to remove any LGA's that were not included in both the 'crime' and 'sport' set since I would not be able to compare the correlation for these given data points.

At this point I created a new dataset telling me how many sports facilities, as well as crimes committed, within each given LGA that both sets included for a given year. The index was the name of the LGA and the columns were called 'crime_by_population' and 'sport_by_population' respectively. I then proceeded to normalize this by dividing the results by the population of the LGA. This was done because different LGA's vary in population and this may well be a large factor in creating differences between the number of crimes and sport facilities between LGA's. I then sorted this new dataset from ascending order based on the size of the normalized crime figure.

Visualization:

From here, I attempted to visualize this result in a few ways to observe any potential trends that may be present between the two variables. Using Matplotlib I came up with a line graph, with the LGA's on the x-axis, and the crime/sport facility rate on the y-axis. Here there appears to be a inverse relation between the two factors as seen in the year 2012 and 2015.

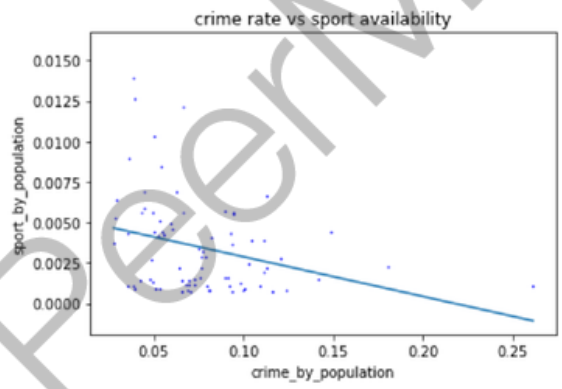


To further understand the level to which there might be a link between the two factors, I also employed the 'Pearson's Correlation'. With the result being -0.31 it is suggested that there is indeed a moderate negative correlation between the two (Cohen,1988), and indicating that a decline in sporting facilities per person does seem to appear as the number of crimes per person increases (this does not imply causality).

	crime_by_population	sport_by_population
crime_by_population	1.000000	-0.306807
sport_by_population	-0.306807	1.000000

(above is table generated when the 'Pearson's correlation' is calculated using python)

Furthermore, I decided to generate a scatter plot between crime rate and relative availability of sports facilities, along with a line of best fit to visualize the general gradient of the trend between the two. With each dot corresponding to an individual LGA, there appears to be a negative gradient (and therefore correlation).



Value of continuation:

This project is pursuable because there is sufficient, reliable data to analyse from and it has not been taken on to an extent that the answer to this question that I am proposing has been answered definitively. In furthering my research under this topic, I will produce results that will either support or reject the notion that increased availability of sport facilities generally lowers the rate of crime in any area. While initial results suggest that there is indeed a correlation between the two factors, what will be interesting is the stance that the data will take upon further investigation, regardless of whether there is seen to be a strong correlation or not.