PeerMark® by Turnitin

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)

"PHASE 2A" BY ANONYMOUS

Does Socioeconomic Status affect choice of sport in Victoria?

The choice of sport varies widely across Australia based on cultural ties to particular states and and cities, i.e Rugby League is very popular in NSW but AFL is the main sporting choice in Victoria. However, within a state, does socioeconomic status have an effect on what sport is prevalent in any particular area?

This question would be of great interest to the government and local councils as it could assist them in building and maintaining sporting complexes for various sports in an effort to grow each individual sport. Local clubs would also be interested as they could change various aspects of their clubs to make the sport more accessible to those who are more willing to play.

This investigation will be innovative as it is not common for reports to be written on why sports are chosen before others apart from safety factors. It brings in another factor as to why certain sports are growing and others aren't, and because of this both sporting clubs and entire sporting bodies can target specific areas to maximize the participation in their sport.

For this investigation I am using a dataset from the state government on sporting and recreational facilities, which lists all sporting facilities in Victoria, where they are located, the number of courts/fields/ovals etc and other information about the facilities which are not relevant to this research question.

The other dataset I am using is the 2015 LGA (Local Government Area) profile, which has hundreds of statistics about each LGA. The statistics that I will be using are median house prices, median income and socioeconomic disadvantage. I will also look at the statistics on how many foreigners are in the LGA and obesity statistics as they could also have an effect on which sports are more commonly chosen and would like to rule out as many other possible factors as possible.

These two datasets can be downloaded at the following URL's:

Sport and Recreational Facilities:

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

LGA Statistics:

https://www2.health.vic.gov.au/about/reporting-planning-data/gis-and-planning-products/geographical-profiles

Processing this data adds a lot of value to the data as before processing occurs, there are hundreds of columns of data from LGA and trying to find meaningful conclusions with data on that magnitude without any processing, visualization or analysis, is almost impossible. Integrating the two datasets together makes it much easier to process, and will lead to getting results that can be visualised in a way that conclusions can be made between the datasets that would otherwise have been impossible.

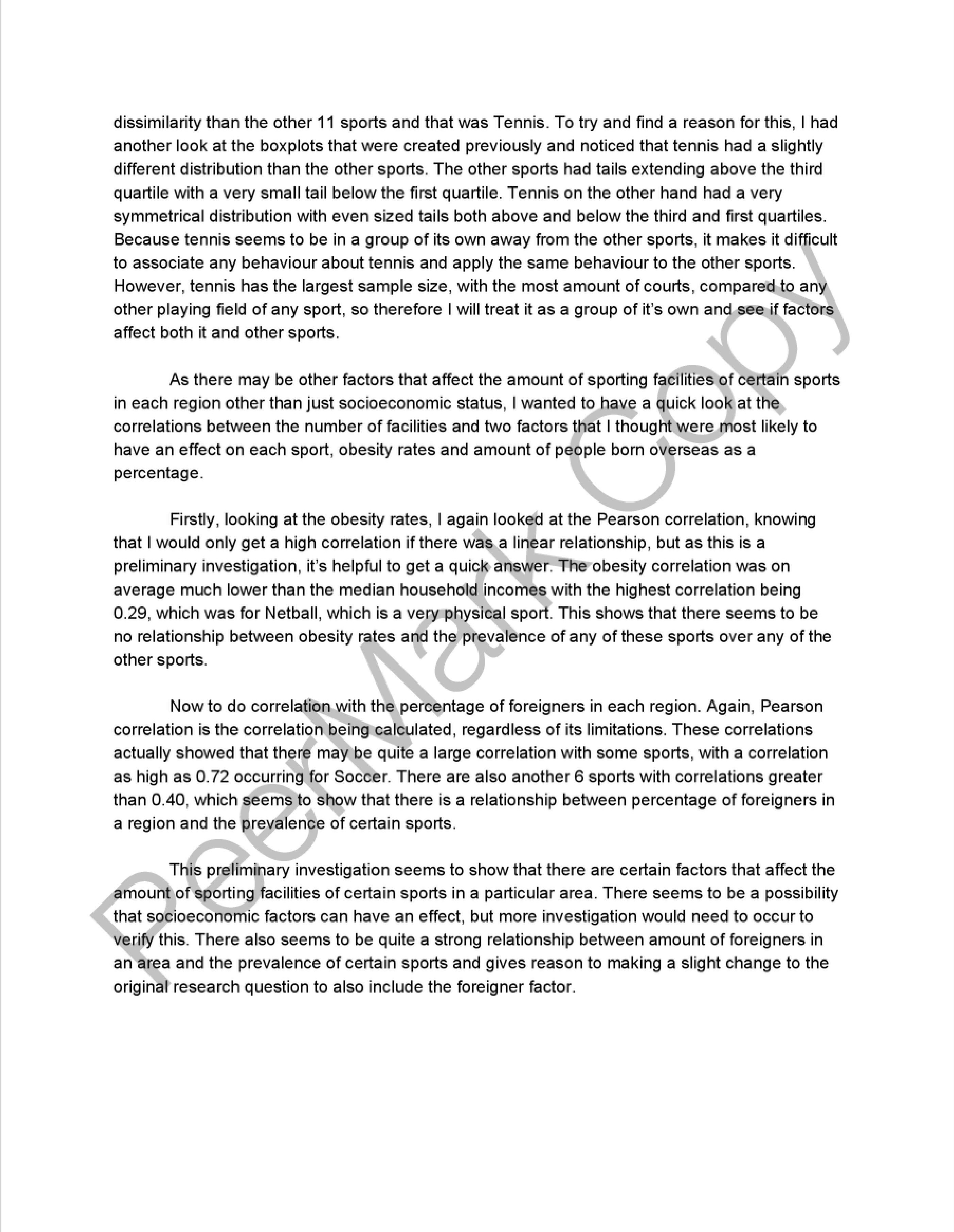
Cleaning the datasets proved to be quite time consuming as the relevant column names had to be found within the 404 columns of the LGA dataset. Once the new dataframe was built, regex operations were required to get the LGA names identical in both the LGA dataset and Sport and recreation dataset. Columns in the LGA dataset consisted mainly of objects and strings and so they needed to be converted to ints to complete operations with the other data. Many sports also had to be deleted from the sporting dataset as there were approximately 50 sports in total and many are not popular sports and wouldn't give a large enough sample size to make any conclusions about them. There were also rows with the same sport but under different names, so some replacing of values was required.

Firstly I ran a boxplot on the number of facilities for each of the 13 sports to look for outliers. Many of the sport had at most 1 or 2 outliers and so there wasn't much concern over the validity of the data, however, possible causes of these outliers would need to be further investigated. There was one sport that the boxplot revealed a lot about and that was rugby league, as it showed that there were only 6 points of data and as such, would provide a poor sample size. For this reason, I decided to exclude Rugby League from any further analysis so as to not draw any false conclusions.

The dataframes were then merged together in an effort to make future work simpler. The first piece of analysis undertaken on both datasets was a pearson correlation between Median Household income and the number of sporting facilities in each LGA. There were correlations as low as 0.023 and 0.062 for swimming and hockey, but 7 out of 12 had correlations between 0.28 and 0.49. These sports were AFL, Soccer, Cricket, Basketball, Softball/Baseball, Rugby Union and Athletics. While these correlations aren't high enough for a relationship to be very likely, but they are high enough to show the possibility that there may be a relationship that needs to be found. The other thing these numbers show is that many of the higher correlations are for the more popular sports in Victoria, ie. AFL, Soccer, Cricket and Basketball. This leads me to believe that maybe this relationship occurs for the large, mainstream sports.

I wanted to see if there were any subgroups within the sports because if there was I could then target separate groups and find characteristics of all the members in the group. To look for this clustering of groups I applied a VAT algorithm to the number of facilities that each sport has.
I found that there were possibly 2 groups, but they were so weak that I can assume that a these two groups have been formed by nothing other than randomness. The fact that there are no prominent groups shows that either all sports have similar behaviour to one another, which would be a good sign for the question as it means that if a variable changes and it affects one sport, it is most likely to affect all of the other sports too.

To check whether the sports were part of one big group or all independant I made a heatmap from all the sports. The heatmap showed that the dissimilarity between the number of facilities that each of the sports has is almost 0. There was one sport that had a much higher



dissimilarity than the other 11 sports and that was Tennis. To try and find a reason for this, I had another look at the boxplots that were created previously and noticed that tennis had a slightly different distribution than the other sports. The other sports had tails extending above the third quartile with a very small tail below the first quartile. Tennis on the other hand had a very symmetrical distribution with even sized tails both above and below the third and first quartiles. Because tennis seems to be in a group of its own away from the other sports, it makes it difficult to associate any behaviour about tennis and apply the same behaviour to the other sports. However, tennis has the largest sample size, with the most amount of courts, compared to any other playing field of any sport, so therefore I will treat it as a group of it's own and see if factors affect both it and other sports.

As there may be other factors that affect the amount of sporting facilities of certain sports in each region other than just socioeconomic status, I wanted to have a quick look at the correlations between the number of facilities and two factors that I thought were most likely to have an effect on each sport, obesity rates and amount of people born overseas as a percentage.

Firstly, looking at the obesity rates, I again looked at the Pearson correlation, knowing that I would only get a high correlation if there was a linear relationship, but as this is a preliminary investigation, it's helpful to get a quick answer. The obesity correlation was on average much lower than the median household incomes with the highest correlation being 0.29, which was for Netball, which is a very physical sport. This shows that there seems to be no relationship between obesity rates and the prevalence of any of these sports over any of the other sports.

Now to do correlation with the percentage of foreigners in each region. Again, Pearson correlation is the correlation being calculated, regardless of its limitations. These correlations actually showed that there may be quite a large correlation with some sports, with a correlation as high as 0.72 occurring for Soccer. There are also another 6 sports with correlations greater than 0.40, which seems to show that there is a relationship between percentage of foreigners in a region and the prevalence of certain sports.

This preliminary investigation seems to show that there are certain factors that affect the amount of sporting facilities of certain sports in a particular area. There seems to be a possibility that socioeconomic factors can have an effect, but more investigation would need to occur to verify this. There also seems to be quite a strong relationship between amount of foreigners in an area and the prevalence of certain sports and gives reason to making a slight change to the original research question to also include the foreigner factor.