Project Report

GitHub URL

https://github.com/DuncanMullaney1997/UCDPA_DuncanMullaney

Abstract

The aim of the following project was to exact meaningful insight from openly available Leinster water polo league fixtures data. The data the contains game details and results for ten teams across multiple seasons spanning five years.

Some goals of the project were to:

A. To import, organise, clean and transform the data to facilitate goals B and C
B. Try to discern the most successful team
C. Try to assess the feasibility of a clustering exercise to try to identify teams based on non-win vs win outcomes.

Introduction

The sport of water polo in Ireland has witnessed tremendous growth in the last number of decades. Whilst yet to boast a professional platform level in the country, amateur water polo is still vibrant and sees healthy participation across most provinces.

Water polo is played at both national (i.e., national league) and regional (e.g., Leinster league) levels.

The data set that I have chosen for this project is Leinster water polo fixture/results listings dating from beginning of June '22, back to beginning 2016/17 season. The data set includes details of all matches played under the auspices of the Leinster water polo official competitions.

Specifically, the data includes, for each game instance:

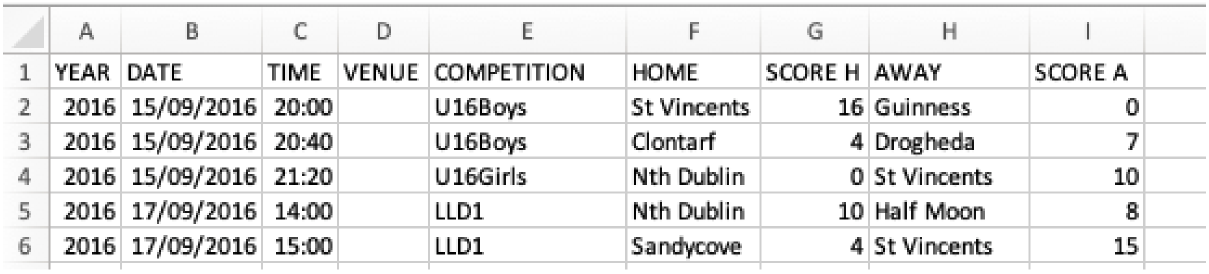
- Game date
- Game time
- Game venue (i.e., the leisure centre/aquatic facility that hosted the game)
- Home team
- Away team
- respective scores/game outcome
- Which category a game was played under (e.g.,
  - U16G being under 16 girls
  - U19B being under 19 boys
  - Ladies being senior ladies
  - Etc.
- Which competition it was played for (e.g., Leinster "Cup" [i.e., a knockout cup] or Leinster "League" [continuous in-season games with an overall winner at season end]).

The data was to be loaded, cleaned, transformed and scrutinized with a view to obtaining some useful/meaningful insight.

Dataset

The data source I chose was "Leinster Water Polo Fixtures/results". The data was accessible on open-source webpages:

- The most recent fixture data (i.e. for the 2021/22 season) was accessible at https://irelandwaterpolo.ie/leinster/. This data was web-scraped, parsed using *read_html* functionality from the python library *beautiful soup*.
- Data for previous seasons, spanning 2019/20 back to 2016/17 was pulled from http://www.leinsterwaterpolo.org/. This data was accessible as downloadable csv files, with a format as described in figure below.



	A	B	C	D	E	F	G	H	I
1	YEAR	DATE	TIME	VENUE	COMPETITION	HOME	SCORE H	AWAY	SCORE A
2	2016	15/09/2016	20:00		U16Boys	St Vincents	16	Guinness	0
3	2016	15/09/2016	20:40		U16Boys	Clontarf	4	Drogheda	7
4	2016	15/09/2016	21:20		U16Girls	Nth Dublin	0	St Vincents	10
5	2016	17/09/2016	14:00		LLD1	Nth Dublin	10	Half Moon	8
6	2016	17/09/2016	15:00		LLD1	Sandycove	4	St Vincents	15

Figure 1 - layout of downloadable csv files from www.leinsterwaterpolo.org.

I decided to choose Leinster water polo data as I play water polo and have participated in the Leinster League and Cup competitions for many years. I am familiar with the teams and the structure of the leagues, so it made an ideal data set to perform an analysis on.

The data contained details of all matches, venues, times, teams, and their scores, as well as both which competition and category the game was under (e.g., U15B Cup, Div2 League, etc…).

The ten teams that participate in Leinster competitions are:

- Drogheda water polo club, abbreviated "DR"
- St. Vincents water polo club, abbreviated "VT"
- North Dublin water polo club, abbreviated "ND"
- Clontarf water polo club, abbreviated "CL"
- Sandycove water polo club, abbreviated "SC"
- Halfmoon water polo club, abbreviated "HM"
- Newry water polo club, abbreviated "NY"
- Guinness water polo club, abbreviated "GS"
- Trinity College Dublin water polo club, abbreviated "TR"
- University College Dublin water polo club, abbreviated "UC"

I was confident that after tidying and cleaning the data, I could use data analytics and some machine learning techniques to gain some useful insight into the data. I was also very interested in this data as it is very relevant to me, particularly as I play for Drogheda, so my team's performance is charted amongst the others for the last 5 years.

Implementation Process

- **Importing data**
  The data was imported from two general sources:
  - Recent fixtures (i.e. for the 21/22 season) was imported from the URL provided in the *Datasets* section of the report. The web-scraping method for importing used was pd.read_html. The python file “process_webscrape.py” achieves this.
  - CSVs for past seasons (2016 – 2020) was imported through the URL provided in the *Datasets* section of this report. The python file “process_csvs.py” achieves this.

- **Cleaning data**
  - CSVs tidying – CSVs had to be formatted in to a uniform way, columns renamed, cleaned and sorted and nulls removed.
  - Recent season tidying – the 2021/2022 data scraped from the web had to tidies also. Regex was used to parse out and normalise the date/time based fields. A “SSN” (Season) variable was then derived from this date field.

- **Joined data together**
  - the two data sets were then joined in to a combined dataset (df_comb)

- **Additional fields**
  - Many additional fields had to be made, including a “CAT” field. This was the breakdown of what the competition was (e.g., U15B denoting under 15 boys category, U19G under 19 girls, etc.).
  - Find and replace method was used to achieve this, taking similar type competitions, and grouping them all together under one general name. this step was taken to simplify the data set.

- **Transforming data**
  - The data had to then be extensively transformed in order to enable graphing (complex data flipping/blending pandas methods such as *melt* and *group by multiple* queries were used extensively, see “process_webscape.py”
  - Functions – two functions were created, used in “apply” methods, to help with transforming the data
    - Panda_strip: remove leading and trailing whitespaces
    - Def_flag: populates a game outcome field per team, with “W”, “L”, or “D” based on game outcomes

- **Graphing methods**
  - I decided that for my analysis, two main charts should be made use of:
    - Stacked bar charts, to display the breakdown of how each team generally performed over each season.
    - Scatter plot, to compare the win vs non-win outcome for each team across all competitions/seasons, with a view to estimating the feasibility of a k-means clustering analysis to try identifying a team solely based on its performance.

Results

A guide to reading the following charts:

Abbreviation	Team
DR	Drogheda
VT	St. Vincents
ND	North Dublin
CL	Clontarf
SC	Sandycove
HM	Halfmoon
NY	Newry
GS	Guinness
TR	Trinity College
UC	University College Dublin

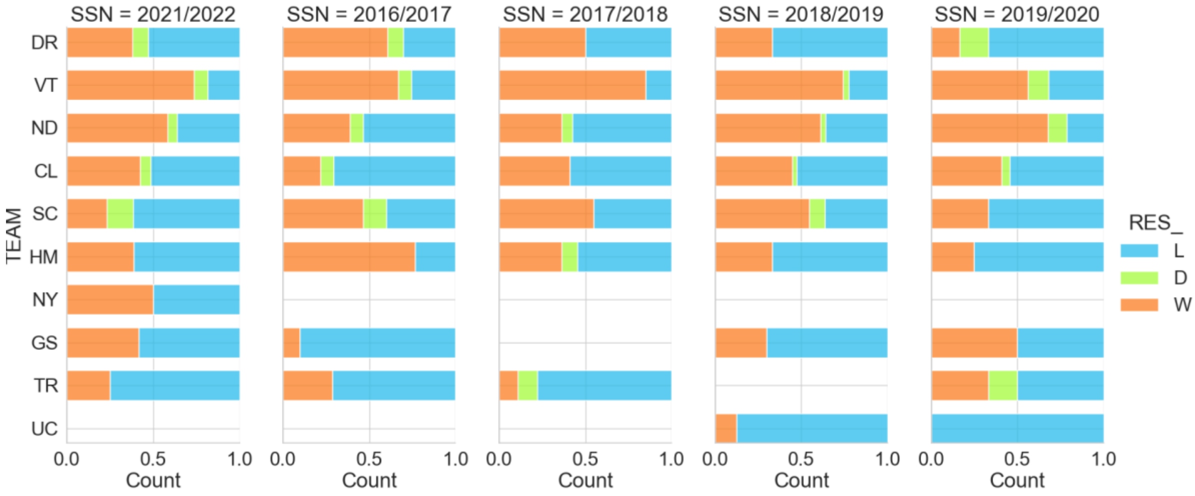


Figure 2 - stacked horizontal bar chart displaying the relative proportion of wins, losses and draws for each Leinster water polo club, and for each year of data.

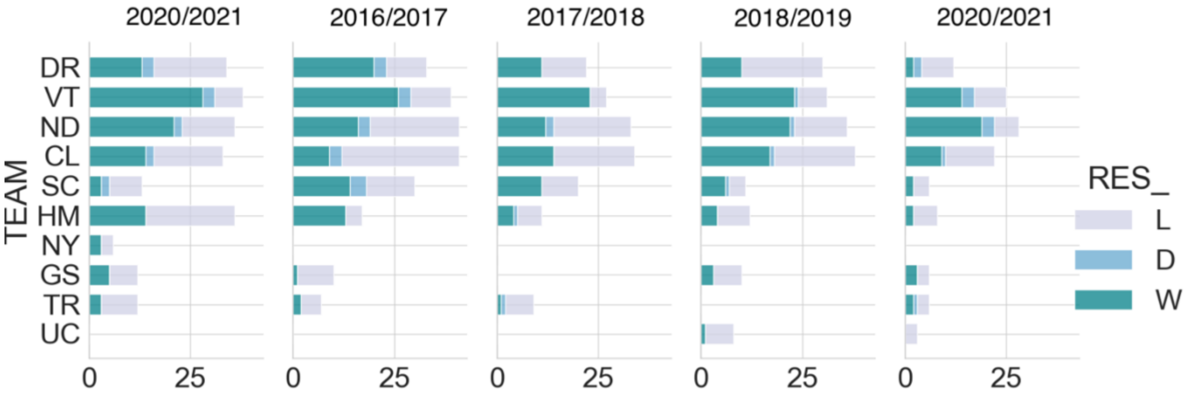


Figure 3 - another stacked bar chart, which shows the same data as in the previous figure, however this shows the actual values instead of proportions.

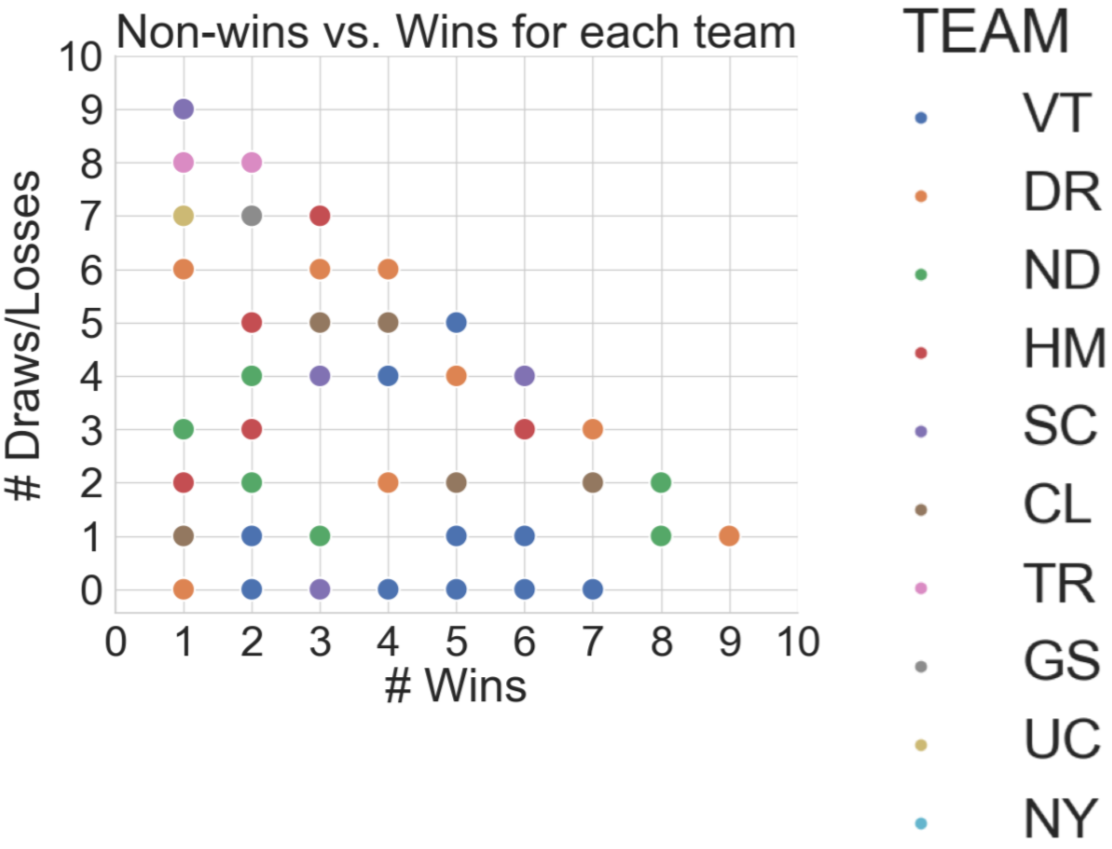


Figure 4 - scatter plot illustrating the number of non-wins (i.e., losses or draws) vs. wins for each team across all categories/competitions.

Insights

- From figure 4, one of the main observations to be made is that “VT”, or “St. Vincents” (an inner-city Dublin team) appears to be one of the strongest teams, with the cluster of blue data points in the bottom right of the scatter plot. This indicates a relatively high win:non-win ratio for them. This suggests that Vincents are a more successful team.

- A cluster analysis on figure 4 chart is certainly feasible and would likely indicate St. Vincents (Blue dots) as a group on their own. I think using more of the data features (i.e., category (U15B, U17G etc…) and a year by year analysis might have offered more insights.
- A sense check on the data was the diagonal slope line that the figure 4 scatter plot makes. This indicates that, overall (i.e., for each competition, for each category and season) each team played no more than 10 games.
- The drop in number of games splayed in the 2019/2020 season is clearly indicated in the right most bar plot.
- The type of analysis conducted here really requires all available features to be used, . Admittedly the dataset did not have many features to work with apart from venue, date, time, and category/competitions. More numerical features would have allowed more machine learning to be conducted.

References

[1] https://pandas.pydata.org/docs/reference/api/pandas.read_excel.html?highlight=read_excel

[2] https://www.crummy.com/software/BeautifulSoup/bs4/doc/