Analysing sports popularity based on Olympic athletes and medals

Group 2 : Eleanor Johnson, Emer McCourt, Georgia Sapsani, Kara Tsonos,
Lauriane Suyin Chalmin-Pui, Rosie Barrie

Introduction **2**
- Aims and Objectives 2
- Background 2
- Report Structure 2

Specifications and Design **3**
- Technical Requirements 3
- Non-Technical Requirements 3
- Project Workflow 3

Implementation and Execution **4**
- Agile Development Approach 4
- Project execution processes 4
- Implementation Challenges 4
- Tools and Libraries 5

Data Collection **5**
- Data Sources 5
- Data Cleaning 5
- API 6
- Limitations 6

Results Reporting **6**
- 1. What are the main patterns in Team GB medals since the modern Olympics were founded (1896)? 6
- 2. Which successful athletes (Team GB and non-GB) can we use to market sports to different generations? 7
- 3. Which new Olympic sports have TeamGB participated in? 8
- 4. Which sports may interest and inspire customers based on Olympic success? 8
- 5. How can this year's Olympic results support up-to-date decision-making? 11

Conclusion **11**

Introduction

Aims and Objectives

The client is a large sports group that includes franchising leisure centres, supporting local councils with their sports programming, and sports marketing for the general UK population. As a business, their vision is to share the passion of taking part in sports activities with everyone and to support access to these sports activities. To achieve this for them, this data science project analyses sports popularity based on Olympic athletes and medals won.

Using historical data on Olympic athletes, the aim is to answer **5 research questions**:

1. What are the main patterns in Team GB medals since the modern Olympics were founded in 1896?
2. Which successful athletes (TeamGB and non-GB) can we use to market sports to different generations?
3. Which new Olympic sports have TeamGB participated in?
4. Which sports may interest and inspire customers based on Olympic success?
5. How can this year's Olympic results support up-to-date decision-making?

Background

The client's business will benefit from an evidence-based approach to understanding the successes of TeamGB and other athletes at the Olympics, especially in terms of prioritising infrastructure, funding, and support for specific sports, and strong marketing to reach target audiences. Olympic success often correlates with broader sports participation and interest. By answering the research questions above, the client can anticipate future demand for certain sports facilities and programmes, identify emerging sports to invest in early and align their long-term strategy with evolving sports preferences. The insights from this analysis could inform several key business decisions:

Facility Planning

- Prioritise development of facilities and equipment for sports showing growing Olympic success and popularity
- Allocate more space and resources to high-performing Olympic sports in leisure centres

Programming and community engagement

- Expand offerings and classes for sports gaining Olympic traction
- Develop targeted programs to nurture talent in sports with TeamGB success
- Organise events and activities centred on popular Olympic sports
- Create youth programs focused on sports with Olympic potential

Marketing and Promotion

- Feature Olympic athletes and medal winners in marketing campaigns
- Highlight Olympic sports where GB performs well to drive interest
- Create campaigns around emerging Olympic sports to build excitement

As a major player in UK sports and leisure, the client's decisions can impact **national sports development**. Using Olympic data to inform their strategy allows them to:

- Contribute to building a stronger pipeline of Olympic talent
- Support sports where the UK has potential to improve Olympic performance
- Boost overall sports participation by promoting Olympic success stories

Report Structure

This report outlines the data analysis undertaken, with sections on the specifications and design, implementation and execution, data collection, and reporting results. It is accompanied by the source data and Jupyter notebooks for all the code.

Specifications and Design

Technical Requirements

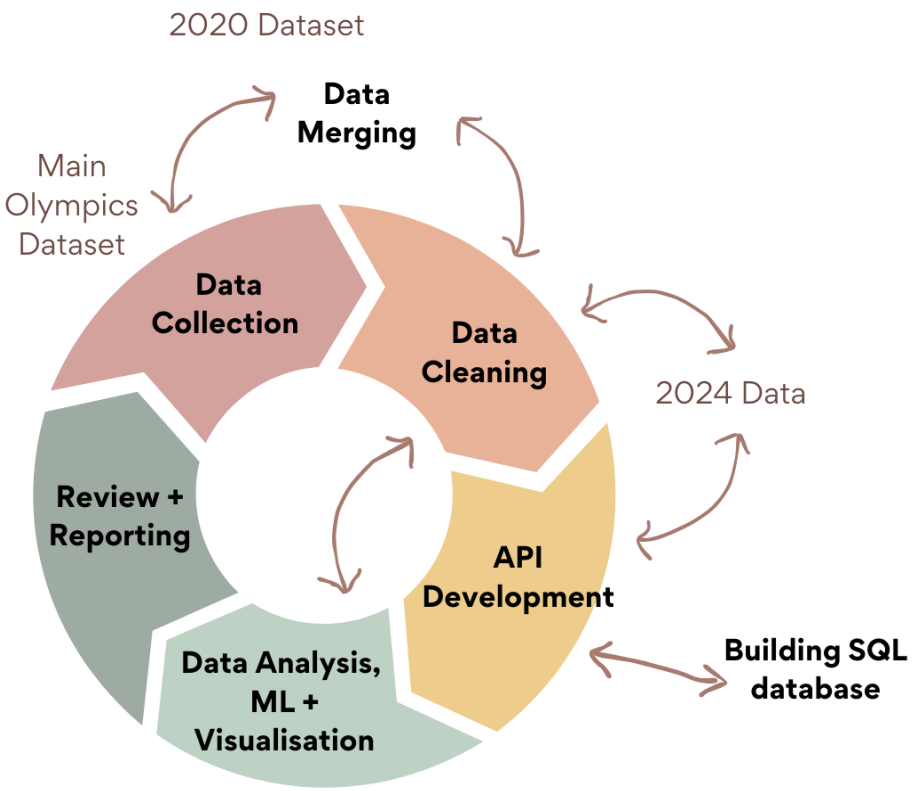
- **Data Sources and Integration:** Complete dataset on Olympic athletes. (In the end, this had to be brought together from 3 different datasets including a historical dataset (spanning 120 years), 2020 dataset and a third data source via an API for 2024 data.
- **Data Handling:** Data cleaning was essential due to missing values, inconsistent formats, and some discrepancies in athlete names. The cleaning process included handling NaN values, removing duplicates, normalising and standardising text data.

Non-Technical Requirements

- **Client Objectives:** The analysis was aimed at helping the client—a large sports group—make data-driven decisions on resource allocation, sports promotion, and strategic planning based on Olympic data.
- **Report Clarity:** The final report needed to be clear and actionable, with insights presented in a way that could be easily understood by non-technical stakeholders. This included visualisations, summary statistics, and clear explanations of findings.
- **Goal Alignment:** The project's goals were aligned with the client's business needs, to identify patterns in Olympic success and potential opportunities to promote sports based on historical and recent data.
- **Outcome Utility:** The project aimed to deliver actionable insights that the client could use to support marketing, facility planning, and community engagement efforts.
- **Team Coordination:** The project required effective collaboration among team members, with regular meetings, clear communication, and shared documentation.
- **Documentation:** Comprehensive documentation was necessary to ensure that all team members could follow the project's progress and understand the analysis process. This included meeting minutes, code documentation, and detailed descriptions of data processing steps.

Project Workflow

The implementation process was iterative and allowed the team to deliver a data analysis project within the deadline in a few weeks.



The requirements of the project were continually clarified with CFG, and thus the project design, and development was being amended. In a similar way, as the team encountered issues with the data, an iterative process to refine research questions, development tools, coding approaches, data management processes to take new ideas on board was also reviewed and re-developed by all team members. Important decisions were agreed by all.

Ultimately, the whole team was on a learning journey not only about data science but about the development processes.

Implementation and Execution

Agile Development Approach

Teamwork was conducted outside of regular lesson times mostly through online meetings and Slack messaging. Each meeting was minuted and included a 'stand up' of what had been done and what any blockers or issues exist. The team tried to plan the week or few days ahead, allocating action points to individuals. It was difficult to estimate time needed given that everyone is still learning. In Slack, the team also shared what had been done, or any review of difficult processes for example, especially if the team member was unable to attend a live meeting due to other commitments or being unwell.

The team did not have allocated 'Agile' roles but all worked together to the best of their ability. Some code review occurred in pairs on specific topics. This flexibility meant that as a whole, the team were able to adapt and take extra responsibilities from other members who were struggling to balance work-life and participate in the course. All members were aware of dates that others were unavailable and able to communicate individual needs and ask for support when needed.

Project execution processes

The project topic was easily decided by all in the team. After initial discussions around mental health in the tech and education sectors, the dataset based on the Olympics was found and all team members were excited to begin the process. Different team members explored different data science tools and approaches, including exploratory data analysis, data cleaning, descriptive statistics, data visualisation, fetching missing data using APIs, creating SQL databases, and machine learning. All team members developed new skills, strengths, and areas of expertise.

Implementation Challenges

Throughout the project, the team faced challenges with the data, with tools, and with remotely distributed teamwork. Indeed these were all threats and weaknesses identified in the initial SWOT analysis. Most of these challenges were overcome together to form the final project, which contributed to the overall learning outcomes of the group. The challenges faced with the data are all detailed in the limitations section below. The biggest issue encountered was missing data. In the limited timeframe of the project, sourcing more reliable and complete data was unfortunately not possible. Being a remotely distributed team, with limited data science experience, knowledge gaps had to be filled and data processes may not always have been the most efficient but they achieved the team's goals.

Ultimately, not all efforts were used or detailed in the final project analysis. Those that are used are detailed in the Data section and those that were developed but not implemented (with the whole team involved in decision-making) are summarised below to provide insight into the team's development, learning, and thinking.

At the outset it was decided to include a SQL Database and explore the potential usage of an API to query the database for the project. The data from clean_olympics_dataset was imported to SQL. Blocks were encountered with the process regarding handling NULL values and Duplicates. Once overcome, the import and querying was successful. An initial test API was created in Python but not completed. Learnings in the process showed that any additional changes in the Jupyter Notebook could be managed in Jupyter and directly imported from the Jupyter Notebook into the SQL database using a SQL connector. It was decided not to proceed down this route by the team.

Tools and Libraries

The team's primary development tools included the following: Jupyter notebook, Collab, PyCharm, MySQL Workbench, and Git. Python libraries used included: pandas, numpy, matplotlib, seaborn, flask, pymysql, scikit, googletrans, datetime, requests, mysql.connector.

Regarding communication and project management, several tools were used. Meetings were held in Google Meets, with frequent communications in a dedicated Slack channel. Meeting minutes were held in Google Docs in a shared Google Drive, which was also used to share files and work collaboratively. A shared Github repository was used to submit the project. Initially, a Trello board was set up to manage the project but this was not used so well, instead relying on Slack and updated meeting minutes as a way of tracking progress. For a small project, this was sufficient.

Data Collection

Data Sources

For this project there were two datasets used, which were combined to give up-to-date data. The first was 120 years of Olympic history and the second was the Tokyo 2020 Olympics Dataset. These two datasets were used together as the first only had data up to the year 2016, as there have been Olympic games since then, the second dataset was found. The joining process is outlined in the *A_Merging.ipynb* Jupyter notebook. This dataset was based on the first, so joining the data together was simple and only involved changing some of the column names to match the existing dataset and switching the athletes names around to match the 'firstname surname' structure of the initial dataset. The switching of the names was done manually in excel as there was a large variety in length of names.

In the process of early analysis, it was discovered that there were some important athletes missing from the dataset. An API was found that could retrieve the data for the top medal winners, which were then manually added to the dataset. This was an important step, as some of the analysis relied on popular athletes and the medals earned.

Data Cleaning

For the full cleaning process, please see *B_Cleaning.ipynb* Jupyter notebook.

The combined dataset initially started with 87,682 rows and 16 columns. The cleaning process involved analysing the uncleaned data to check the datatype and to see how many NaN values were present. The data was also sorted to make further cleaning and the future analysis clearer.

Once this initial analysis was completed, the cleaning process began by replacing the NaN values in 'Age' and 'Medals' with the median for 'Age' and the string 'None' for 'Medals'. These columns were likely to be useful in the analysis, so by replacing the NaN values, the data would continue to be relevant. Other columns that had NaN values, but did not relate to the analysis, were removed. Duplicate rows were also removed. A new column called 'Country' was added, as the team name did not always relate to the NOC (National Olympic Committee) . This meant that during analysis, it was clear which country was being referred to.

In the final stages of cleaning, the data was all turned to uppercase and the whitespace was removed. This meant it was clearer to read, as the data was in various different forms. During the analysis, there was a discovery that there were two very similar sports: 'Equestrian' and 'Equestrianism'. From researching them, they proved to mean the same thing so the sport was renamed 'Equestrian'. The final file contains all data, as this is also used in the analysis, which post cleaning has 87,295 rows and 13 columns.

API

Using the API detailed in the *API.ipynb* Jupyter notebook, data was collected on the Paris 2024 Olympic Games. This included information about the Country rank, medal count for gold, silver, bronze and total medals won. The top ten ranked countries who won the most medals at the 2024 olympics were retrieved from the API.

Another API from the same source was able to find the sport for which medals Team GB won in the Paris Olympics . This evolved into retrieving the day the medal was won, exactly what medal was won, and the sport it related to. Due to the necessity of filtering through over 4000+ entries spread across multiple pages, this operation has a 4 minute runtime to find Team GB information. Once the API had run, the retrieved information would be stored in the *olympics.sql* database within the medal_events table which can then be accessed on the local server using Flask API design.

Limitations

Throughout the analysis, several limitations were discovered and noted. The first limitation is the source of the data - a website where any individual can upload datasets without having evidence to prove their accuracy. If this report and analysis were to be used in a business, the data collection would be done in a different manner, using reliable sources. The next limitation discovered was the data itself. As it only went up to the year 2016, the data was not as recent as initially hoped. To navigate this limitation, a more up-to-date dataset with data from the 2020 Olympics was sourced and merged with the existing dataset and an API was found to gather data from the 2024 Olympics.

Another limitation noted was the discovery of missing athletes. During the initial analysis phase, it was discovered that some notable athletes were missing from the dataset. As these athletes would be relevant to the final analysis, data about athletes and their medal count was gathered using an API, and then each athlete was individually researched to add the correct information to the dataset. Although this process was time-consuming, it was important for making the analysis as accurate and useful as possible.

Another limitation after cross-referencing medals won from the API against the Olympics official website, it was noticed that four medals were missing (boxing x1 bronze, track cycling x2 bronze, athletics x1 silver), this caused some limitations with understanding the reasons as to why these were missing from the API data, however, the missing medals were identified and manually added to the medal results CSV file.

The final limitation discovered was during the analysis process. It was found that there was a discrepancy with the number of medals received in 2020 for Great Britain, and what the actual count was, which was much lower. After discovering this, it was found that the count was per person who received a medal, rather than per event, which caused some of the data to inflate. The missing athletes mentioned previously, also had an impact on this data. Some limitations in the clustering and regression analysis are due to the missing data and that a more complex Machine Learning model needs generating to investigate, with better clarity, the sports that will be popular in the future based on a bigger number of participants.

Results Reporting

1. *What are the main patterns in Team GB medals since the modern Olympics were founded (1896)?*

As demonstrated in Jupyter notebook *1_Medal_Patterns.ipynb*, Team GB has been awarded a total of 784 medals in the Olympic games; 749 from the Summer Games and 35 from the Winter games. The most successful year overall for the British team was in 1908 with a total of 101 medals but in recent years (post-1945) the most successful year for Team GB was the 2016 Summer Games with 52 medals. Team GB has achieved only 1 medal in several Winter Games (1952-2002) but this has only occurred once in the Summer Games in 1904.

Since 1996, Team GB have seen a notable increase in the number of medals achieved and the proportion of silver and gold medals won has increased in comparison to earlier years. This improvement could be linked to the introduction of National Lottery funding in 1997 for athletes and their training.

The majority of medals achieved by Team GB are in athletics, swimming, rowing, hockey and cycling. In the Winter Games specifically, they have won medals in ice hockey, figure skating, bobsleigh, alpinism, curling and skeleton. Alpinism in particular is interesting as the medals were awarded during the closing ceremony of the 1924 Olympic Games to individuals who had participated in an unsuccessful expedition of Mount Everest in 1922.

The youngest athlete to win a medal for Team GB was 13 years old in the 2020 Olympic Games and won a bronze medal for skateboarding. The oldest British athlete was 73 in the 1948 Olympic Games and won a silver medal in painting and graphic arts.

2. *Which successful athletes (Team GB and non-GB) can we use to market sports to different generations?*

It was found in Jupyter notebook *1_Medal_Patterns.ipynb* that the most decorated Team GB athletes typically win medals in swimming or cycling events. Further exploration in *2_'Athlete_Marketing.ipynb* found that successful athletes outside of swimming and cycling include Jack Beresford (rowing), Ben Ainslie (sailing), Seb Cole (athletics). The top female athletes are Rebecca Adlington (swimming), Margaret Cooper (swimming), Charlotte Dujardin (equestrian) who have all earned 4 medals in their respective sports.

Name	Sport	Medals
Duncan Scott	Swimming	10
Jason Kenny	Cycling	9
Adam Peaty	Swimming	8
James Guy	Swimming	8
Bradley Wiggins	Cycling	8

In *1_Medal_Patterns.ipynb*, the top performing sports for Team GB were also discovered. By looking at successful athletes from those sports, it can be determined who could potentially appeal to customers. While the top three athletes in cycling are from Team GB, the fourth is Burton Cecil Downing who competed for the USA in 1904 with 6 medals, and the fifth is Jens Fielder from Germany with 5 medals.

Name	Country	Sport	Medals
Allyson Felix	USA	Athletics	13
Michael Phelps	USA	Swimming	28
Elisabeta Lipă	Romania	Rowing	8
Eva Roma Maria de Goede	Netherlands	Hockey	4
Jason Kenny	Great Britain	Cycling	9

Outside of the aforementioned sports, other potential athletes are Larisa Latynina (gymnastics), Sven Fischer (biathlon), Marit Bjørgen (cross-country skiing) and Edoardo Mangiarotti (fencing).

In terms of generational appeal, the majority of successful athletes have achieved success in more recent years. However, by dividing the years into separate groups, it can be seen which athletes could appeal to individuals of different ages by looking at athletes of a similar age or who were successful when our customers were younger.

Age Group	Years	Potential Athletes
< 11	2013-present	Emma McKeon, Andre de Grasse, Caeleb Dressel, Penny Oleksiak, Duncan Scott
12 - 27	1997-2012	Michael Phelps, Inge de Bruijn, Ian Thorpe, Natalie Ann Coughlin
28 - 43	1981-1996	Krisztina Egerszegi, Manuela di Centa, Matt Biondi, Vitali Scherba
44 - 59	1965-1980	Nikolai Andiranov, Sawao Katō, Mark Spitz, Alexander Dityatin
60 - 78	1946 - 1964	Larisa Latynina, Takashi Ono, Boris Shakhlin, Edoardo Mangiarotti
79+	< 1946	Giulio Gaudini, Gustavo Marzi, Eugen Mack, Heikki Savolainen

3. *Which new Olympic sports have TeamGB participated in?*

The sports added in 2020 include Skateboarding, Sport Climbing, BMX racing, BMX freestyle, and surfing. This means there has been more mediatisation of these sports and it would be useful to capitalise on this exposure through targeted marketing.

As detailed in Jupyter notebook *3_NewSports.ipynb*, the demographics of athletes participating and winning in these sports are very young. In women's skateboarding in particular, the youngest athlete has been 12 years old. The table below summarises some demographics to target as potential customers for club participation, marketing, merchandise. In addition, because these sports can be practised and are likely to be popular in dense, urban areas, the suggested target is young urban dwellers for these sports. Surfing would only be accessible in coastal areas such as Cornwall and Wales.

Sport	**Target demographic**
Skateboarding	Young teens, including young girls
BMX racing and freestyle	Older teens and young adults
Sport climbing	Older teens and young adults
Surfing	Adults in 20s-30s living in coastal areas

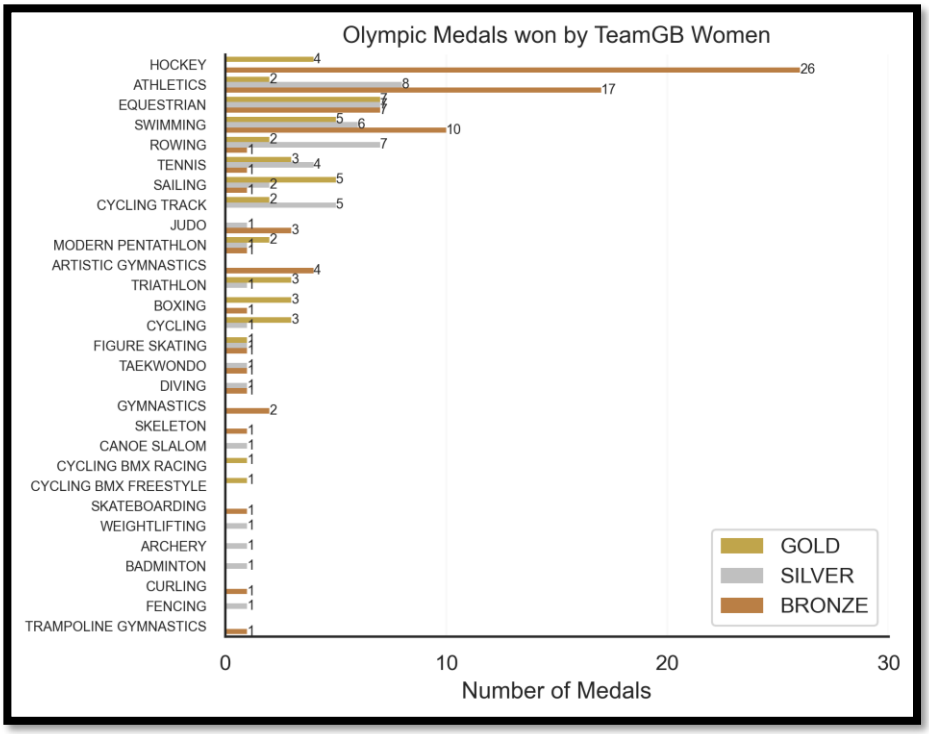
The data has also been queried to see whether there are any older Olympic sports that didn't get immediate TeamGB participation or qualification. The following sports met this criteria: cross country skiing, ski jumping, biathlon, judo, luge, volleyball, handball, rhythmic gymnastics, short track speed skating, and golf. These sports could perhaps use a boost in terms of popular image and participation base. Volleyball, in particular, has only ever had 1 TeamGB qualification in 2012 in London and no medals. A potential business decision could be to support existing volleyball clubs to expand their networks, to make links with international volleyball players, and to cultivate future champions with hopes of a first ever volleyball medal for Great Britain.

4. *Which sports may interest and inspire customers based on Olympic success?*

The Olympic Games are highly mediatised events. Olympic medallists often become well-known sportspeople, especially those who go on to win multiple medals across events or years. When TeamGB are in a medal event (after heats, for example), these are usually broadcast in Great Britain. These sports therefore get more exposure both during and after the event. We have used the data to analyse which sports disciplines are likely to interest and inspire our customers based on past Olympic success.

As detailed and visualised in Jupyter notebook *4a_Sports.ipynb*, the 10 sports in which Team GB has the most all time medals are the following: athletics (111), rowing (91), swimming (70), hockey (67), cycling (65), sailing (56), equestrian (35), tennis (29), shooting (24), and boxing (21).

Women in the UK are less likely to take part in sport than men (Sport England, 2024) so the decision was made to focus on which sports could inspire women based on past Olympic success. As seen in the graph below, the top sports in which Team GB women have won the most all time medals are the following: hockey (30), athletics (27), swimming (21), equestrian (21), rowing (10), sailing (8), tennis (8) and cycling track (7). These sports would be good target sports for encouraging women into amateur sport.

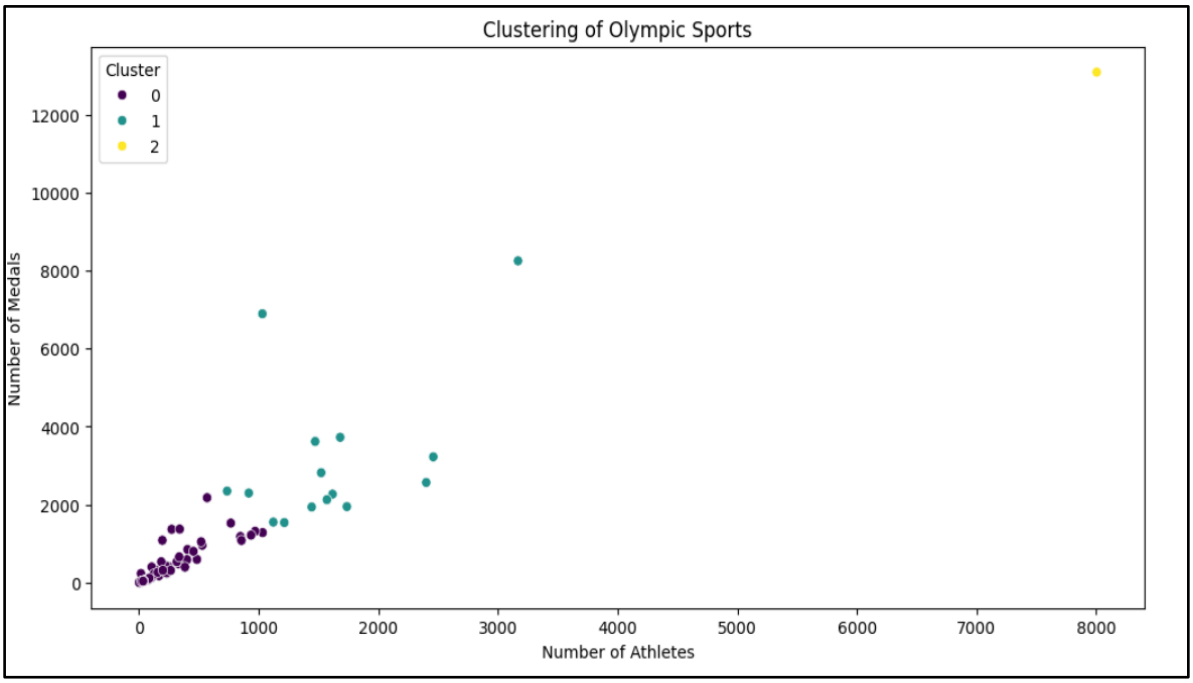


Which sports does team GB have a medal but no gold medal?

There is also an interesting category of sports whereby TeamGB has won Bronze or Silver medals but never a Gold medal. This idea could be marketed as a search for Britain's first gold medallist in the sport! As found in the Jupyter Notebook *4a_Sports.ipynb*, GB has yet to win a gold medal in the following sports despite other podium successes: Badminton, Judo, Lacrosse, Rugby (Union), Rugby Sevens, Skateboarding, Taekwondo, Trampoline Gymnastics.

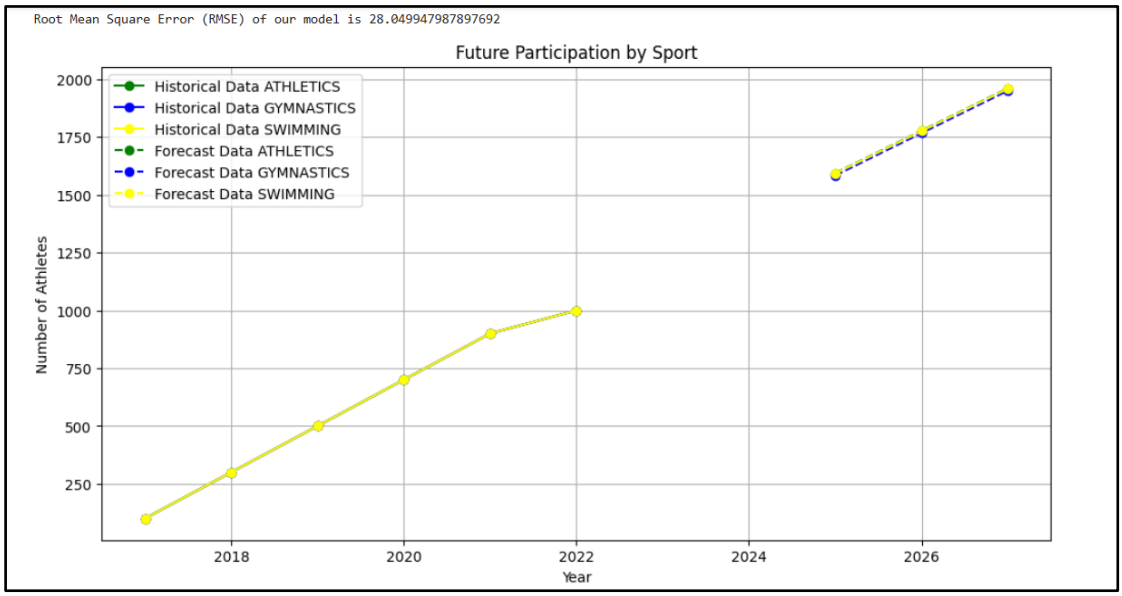
Which sports will be popular in the future?

Using the Machine Learning (ML) library Scikit in the Jupyter notebook *4b_ML.ipynb*, clustering was used to sort data into groups and also create an ML model to forecast the sports that will be popular in the future to support the decisions on the leisure centre's future. Looking at the number of teams participating by sport, 3 sports that have more participants than others: *Athletics, Swimming, Gymnastics*.



Performing predictive analytics by using clustering, the data was split into 3 groups. Each dot represents a different Olympic Sport.The graph below identifies the 3 different clusters. Cluster 0 has a lower number of athletes and medals. Cluster 1 has a moderate number of athletes and a higher number of medals. Cluster 2 is an outlier, which represents a sport that has an exceptionally high number of athletes and medals.

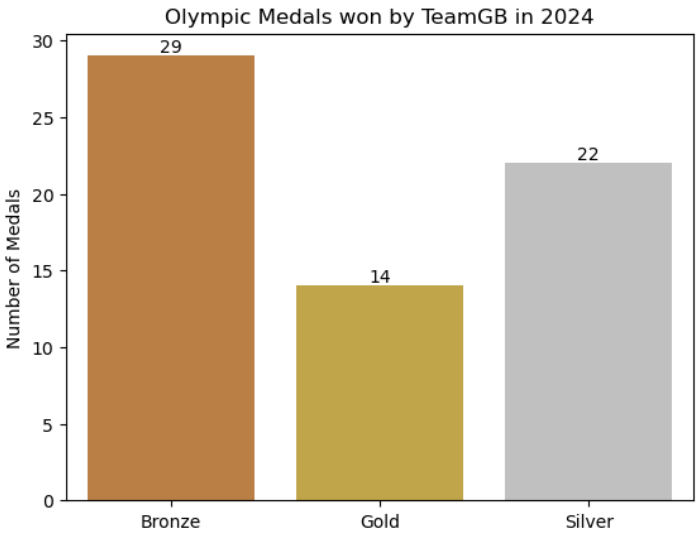
Regression analysis for the three more popular sports helps us understand and forecast their future. The RMSE is 28 which means that on average, the difference between the number of participants predicted by the model and the actual number of participants is about 28. The graph below shows that participation in all three sports (Athletics, Gymnastics, Swimming) has been increasing. The dashed lines represent the predicted trend that more athletes will be participating in these sports in the future.



5. *How can this year's Olympic results support up-to-date decision-making?*

This question is answered using code in Jupyter notebook *5_2024results.ipynb*, which is based on data fetched through an API described in the API section.

TeamGB won a total of 65 medals. The 14 Gold medals were won in the following sports: Equestrian, Cycling Mountain Bike, Triathlon, Shooting, Swimming, Rowing, Trampoline Gymnastics, Cycling Track, Athletics, Sailing, and Sport Climbing. These are therefore also sports that could be used with very contemporary marketing campaigns and those which we can expect for there to be a surge in demand in the coming season.



Across all medals, sports in which multiple medals were won include athletics (10), cycling track (8), rowing (8), swimming (5), diving (5), equestrian (5), canoe slalom (4), triathlon (3), artistic gymnastics (2), sailing (2) and shooting (2) so these are also sports that will likely benefit from a boost in confidence about the abilities of the entire sporting team (athletes, coaches, facilities, investment, etc. It is likely to see continued interest in these sports in future years, with chances of more (gold) medals. If submitting funding applications with local councils, these are sports that are likely to be funded for increased local facilities and staffing increases.

Conclusion

This research has provided valuable insights into the patterns of Team GB's Olympic success and has identified key trends in sports popularity that can inform strategic decisions for the leisure centre.

Team GB athletes such as Duncan Scott, Bradley Wiggins and Adam Peaty are recognisable names for the British public and their respective sports could be marketed well to consumers. Younger consumers may have a stronger interest in athletes such as Emma McKeon and Michael Phelps whereas older individuals may recognise names such as Larisa Latynina and Edoardo Mangiarotti. The introduction of new sports (skateboarding, BMX, surfing, sport climbing) presents the opportunity to engage a new generation and motivate others. Both clustering and regression analysis helped to identify sports (athletics, gymnastics, swimming) with the highest participation and also forecast their popularity.

Overall, this project has laid a strong foundation for making informed decisions that will not only benefit the business but also contribute to promoting a healthier and more active population.