Olympic Swimming and Running Trends

CSC-475: Seminar in Computer Science
Fall 2023

Grace Houser
grace.houser@furman.edu
Department of Computer Science
Furman University
`grace.houser@furman.edu`

Abstract

Swimming and running have been an Olympic sport since the first modern Olympic Games of 1896. Since then, both sports have evolved over time with changes in rules, techniques, nutritional information, and overall improvement of athletes over time. This study takes a look at the trends of swimming results over the years of the Olympic Games since 1912, which is the earliest Olympic Games that has the needed recorded information. Additionally, this study compares the Olympic swimming and running results of the same distances over time. In conducting this analysis, the library Matplotlib in Python was used to perform data visualization. Results from the visualizations showed that women's swimming trend lines tended to be below or on top of running trend lines, in comparing the two sports. The opposite was found for men's trend lines. This is suggested to be a result of the fact that women are better marathon swimmers than men due to biological differences.

1 Introduction

The sports of both swimming and track and field have been some of the oldest sports competed in. For swimming, the sport has consistently been a favorite one to watch in the Olympics. In 2000, swimming was the second favorite sport in the Olympics with 18% favored responses, only 3 percentage points below the favorite sport, track and field with 22% [1]. For the next Olympic year in 2004, swimming and track and field tied for the favorite Olympic event with 21% [1]. Afterward, the 2008 Olympic watchers favored swimming with 27%, before favoring track and field with 18% [1]. Perhaps part of this interest had to do with the incredible achievements of Micheal Phelps. In 2000, Michael Phelps competed at the age of 15, which was the youngest Olympic athlete on Team USA in 68 years. Then, in the 2008 Olympics, Phelps won eight gold medals—one in every event he raced [2]. Some of this success is accredited to a new swimsuit invention that reduced skin friction drag by 24%; however, Phelps continuing winning gold medals in the 2012 and 2016 Olympics, despite the suit getting banned [3]. Regardless, with athletes like Phelps in the swimming community, the achievements in which elite swimmers perform constantly outdo each other. This leads to the question: how much do swimming results change over time? Is there a trend in swimming results over time?

Furthermore, this paper will explore the comparison between Olympic swimming and running events. Because swimming and running are both racing events, it will be interesting to see if the rate of progress moves in similar ways amongst a variety of distances. Especially since the Olympics consist of elite athletes, it is interesting to wonder how the different sports compare. With this interest in mind, the hope in this study is to find information that tells us the similarity in competitiveness between the two sports, despite the different means of racing: by land or by water. In order to dive into the question of swimming results over time and investigating the comparison between swimming and running results over time, this study takes a data analytics approach. First, data will be found and processed. Then, results will be found using Python code and visualization libraries.

2 Related Work

In today's age, data availability of competition times can be found practically instantaneously. As a result, understanding the "nature and prevalence" of competitions, particularly in the Olympics, has been of renewed interest [4]. Before the ready access to data, the history of swimming research can be dated back to 1538 when Colymbetes—the first book dedicated to swimming—was published by Nikolaus Wynmann [5]. Since then, various studies of swimming techniques have been observed, tested, and implemented in the sport. Recently, stroke and suit analysis has been of particular interest in swimming analysis. Stroke analysis is important because arms direct the swimmer through the water when swimming. For example, stroke length and stroke frequency are closely looked at when analyzing efficiency [6]. As for suit analysis, the inspiration behind such research stems from the 2008 Beijing Olympics. At this Olympic Games, swimmers performed faster than prediction models estimated, and this was due to a new swimsuit invention that used a special fabric that changed body buoyancy [7]. Even the Japanese coach broke a sponsorship because of the advantage of the swimsuit [3].

The history of running research, however, is harder to track. Part of this is because running is inherent to human life. In pre-historic times, running was essential for humans to escape life-or-death situations, like outrunning a predator [8]. It was not until 776 BCE that the first official running competition was recorded in the first Olympics held in Greece [8]. After this, running techniques were surely observed; however, it is difficult to find any first recordings of such analysis. Today, as we have technical advantages that can gather and manipulate large amounts of data, questions around limits of world records arise. For example, one study found that, based on given data, there is no limit to human performance in the near future [9].

Beyond the similarity of swimming and running in that they are both racing events that don't use any props other than the human body, it is interesting to note the popularity of these sports in the Olympics [1]. In 2000, track and field was America's favorite sport in the Olympics with 22% responses in a survey given by Gallup Poll, and swimming was the second favorite sport in the Olympics with 18% responses [1]. For the next Olympic year in 2004, track and field and swimming tied for the favorite Olympic event, both receiving 21% of responses. Afterwards, survey respondents for the 2008 Olympics favored swimming with 27%, before favoring track and field with 18%.

Some of the dedication and popularity of running and swimming may stem from the fact that the sports are some of the earliest practiced and competed. For example, in the very first Olympics in 776 BC, a "footrace," or running event, was the only sport to take place [10]. Then, swimming officially became a competitive sport with its first championship meet held in 1846 in London [11]. By the first modern Olympic Games in 1896, swimming made its debut as an Olympic sport.

3 Data

For this project, two datasets were outsourced: one for swimming and one for running. The dataset used for swimming data is "Olympic Swimming History 1912 to 2020" which was found on the website Kaggle [12]. Kaggle is a platform where users can both publish and find free datasets from any variety of topics. In the case of this project, Kaggle held the "Olympic Swimming History 1912 to 2020" dataset that was used for this project. This dataset was uploaded by Data Science Donut, who is classified as a Datasets Expert by Kaggle. To collect the data, Data Science Donut outsourced information from the official Olympic Games website. Additionally, this dataset was updated in April of 2023 by Data Science Donut.

For running data, the dataset used is "Olympic Track and Field Results" which was also found on Kaggle [13]. This dataset was uploaded by Jay Ravaliya, who is both a Software Engineering at Apple and an active member on Kaggle. Jay also scraped data from the official Olympic Games website and last updated this dataset in 2016. However, once getting this dataset, all instances of field events were dropped because their is no swimming event that is comparable to a field event.

As the title of this dataset suggests, "Olympic Swimming History 1912 to 2020" includes all known Olympic swimming result times from every operating Olympic Games from 1912 to 2020. Additionally, the running dataset was renamed to "Olympic Running History 1896 to 2016" for consistency, as it also includes all known Olympic running result times from every operating Olympic Games from 1896 to 2016. In terms of instances, the only notable difference between the two datasets is that the swimming dataset includes the top eight times of each event, and the running dataset includes the top three times of each event. It is additionally important to note that because the Olympic Games were canceled due to World War I, there is no data from the year 1916. Additionally, there is no data from the years 1940 and 1944 because the Olympic Games were canceled due to World War II. However, because we have 28 Olympic years of data, these three years of no information should not hinder any trend lines that we may see in the data.

See Table 1 for a table of dataset attributes that are discussed in the following paragraphs.

Swimming	**Running**
Location	Location
Year	Year
Athlete	Name
Gender	Gender
Team	Nationality
Rank	Medal
Distance	Event
Stroke	Result
Relay?	
Results	

Table 1: Olympic swimming and running dataset attributes

In order to understand the meaning of the attributes in "Olympic Swimming History 1912 to 2020," it is important to understand swimming as a sport. In the sport of swimming, competitions are called "meets" and are organized similarly to running races. For a certain event, eight people swim simultaneously because, in a standard Olympic-sized pool, there are ten lanes. Swimmers compete in the middle lanes, allowing a space for the waves to crash against the wall and not disrupt their movement through the water. Each time the eight swimmers swim their event simultaneously, the race is called a "heat."

There are different types of swimming events that can be competed, and events vary in two aspects: distance and stroke. In an Olympic-sized pool, one length is 50 meters long. However, swimmers may swim in events with distances of 50, 100, 200, 400, 800, or 1500 meters. As far as "strokes" go, there are four different types: freestyle, backstroke, breaststroke, and butterfly. Each different stroke of swimming can be thought of as a different style. For each style, there are unique rules that make the stroke complex in its own way. By combining a distance and a stroke, a swimming event is made. However, it is important to note that not all combinations of distances and strokes are an Olympic event. One reason for this is since some strokes are more difficult than others, having longer distances of difficult strokes becomes extremely strenuous very quickly. Consistent distances between strokes, however, is the 100m and 200m event. Therefore, in the analysis, the swimming trends between the 100m and 200m events will be compared, since both distances are a commonality between all strokes.

Furthermore, for the swimming dataset, under the "Stroke" attribute, there is an "Individual medley" and "medley" option in addition to the four strokes. An "Individual medley" is an event where each individual swimmer swims all four strokes in one race. A "medley" is a relay event where, in one heat and in one lane, a team of four people swim one of the four strokes right after one another. Under the "Distance" attribute, one may notice that a medley is labeled as "4x100." This simply means that four people swim 100m distance each. Thus, anytime "4x100" appears, we can assume that it is a relay event. Our knowledge can be confirmed with the binary attribute "Relay?" that holds a 0 if the instance was not a relay and a 1 if the instance was a relay.

As for the running dataset, the "Event" attribute both included distance and style. For example, the format of the inputed data is "*Distance*M *Style* *Gender*", where M stands for meters. The different styles may include hurdles or relays, for example; however, if the event was a simple running event, the style part of inputed data was excluded.

In total, the swimming dataset has ten attributes and 4,360 instances. Each instance corresponds to an individual person's or team's race. Additional attributes that were not discussed above include the location the swimmer performed, the year the swimmer performed, the gender of the swimmer, the country the swimmer represented, the name of the swimmer, the time the swimmer performed, and the rank the swimmer received. The running dataset consists of eight attributes and 1,560 instances.

Similarly to swimming, each instance corresponds to an individual person's or team's race. Additional attributes that were not discussed above include the location the runner performed, the year the runner performed, the gender of the runner, the nationality the runner represented, the name of the runner, the time the runner performed, and the medal the runner received.

Finally, it is important to note that when a swimmer is disqualified from an event, their "Results" input is either blank or some other text string indicating their disqualification. For example, "Disqualified" or "Did not start" indicates a disqualified person for that given event. In the swimming dataset, because only eight disqualified instances were found, they were removed. This left the swimming dataset with 4,352 instances, which will not significantly change any patterns in trend lines. As for the running dataset, under "Result," if a time was not present, then "None" would show. There was a total of 37 instances found with "None," and they all had to be removed since the information could not be outsourced elsewhere. This left the running dataset with 1,523 instances. However, after further data processing that will be further explained in "Methods," this paper's results are only effected by 15 instances, which in total should not significantly hinder trending lines.

4 Methods

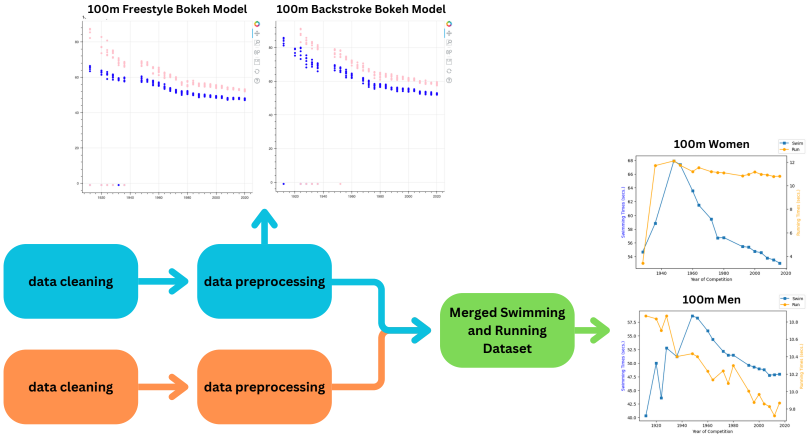

Figure 1: Pipeline of this project

After finding the "Olympic Swimming History 1912 to 2020" dataset, initial data cleaning needed to be completed. First, eight instances had to be dropped because the given swimmers were disqualified or had a blank time entered under the "Results" attribute. These instances only accounted for 0.18% of the dataset and were found over five different Olympic years, so the overall data should not be misrepresented by such data deletion. Furthermore, the "Distance (in meters)" and "Results" attributes were implemented as strings in the original dataset, but needed to be recognized as numbers so they could be numerically manipulated and analyzed. These changes were done in the coding language, Python, and implemented within Google Colab. Google Colab is a service that allows collaboration in Python-based projects, similarly to how Google Docs can be used for English papers.

Once data cleaning was finished with the swimming dataset, preliminary findings were computed with Bokeh, which is a Python library. Bokeh is helpful for data visualization, so in using it for preliminary use, it was beneficial to see the beginning trends of the swimming dataset. Additionally, Bokeh was used because it has more interactive features. Since the intent of the findings are to be published on GitHub, using Bokeh would allow for a better user experience. For example, one interactive feature of Bokeh is that the viewer can choose the distance and stoke (i.e. event) that visualization they want to see.

After preliminary visualization was done, the "Olympic Track and Field Results" dataset was sourced and initial data cleaning also needed to be completed. First, field events had to be dropped from the dataset because they are not races, and therefore, are not comparable to swimming, as swimming is a racing event. Hurdle track events were also deleted because there is no comparable swimming event

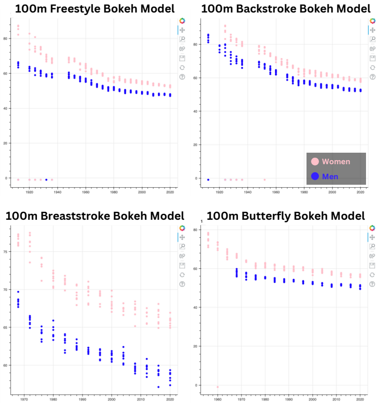

Figure 2: Comparing women and men's 100m swim results in different strokes

that uses an obstacle prop. Additional track events deleted included any distance that is not competed in swimming, like the marathon. Furthermore, the attribute "Results" had to be standardized because in the original dataset, the attribute was a text string input and could not be manipulated numerically.

Then, because the end goal is to compare both swimming and running Olympic results over time, data preprocessing had to be completed before the datasets could be merged into one. Merging both datasets is beneficial because it allows a graph to work with just one dataset, instead of overlaying two datasets of potential different formats. First, it was decided that freestyle would be the stroke of choice that is looked at for swimming. Because freestyle is both the most basic stroke and has the most variety of distances, it was chosen to maximize the amount of potential comparisons with the running dataset. Then, since the swimming dataset has six instances per event and the running dataset has three instances per event, each event per year per dataset was averaged. Therefore, average Olympic results for swimming and running could be compared per event over the years. Between the standardizing of the attributes and the averaging of Olympic results, both datasets could be merged, and with the use of Pandas library in Python code, merging the datasets was implemented.

Finally, the Matplotlib Python library was used to create visualizations of the comparison between average results of Olympic swimming and running events. The swimming trend line is represented in blue and the running trend line is represented in orange. Additionally, the swimming y-axis is on the left and the running y-axis is on the right. It is important to note that the scaling between the swimming y-axis and the running y-axis are different. This was done so that the graphs could overlap eachother. Otherwise, because running is much faster for humans to do than swimming is, the running trend line would be largely below the swimming trend line. Two examples of the findings is included in the project pipeline, but is further discussed in the Results section.

5 Results

Comparable graphs between swimming and running included the women's 100m, men's 100m, women's 200m, men's 200m, women's 400m, men's 400m, women's 800 m, men's 15000m, women's 100m relay, and men's 100m men's relay. Men's 800 meter could not be compared between sports because there is not enough data for swimming's 800 meter men's freestyle race. This is because Tokyo's 2020 Olympics was the first time since the 1904 Olympics that men swam the 800m freestyle race. So, not only are there 29 Olympic years of data missing, but also, the running dataset only includes data up to the year 2016; thus, men's 800 meter Olympic race could not be compared between swimming and running. Additionally, women's 1500 meter race also could not be compared between sports because there is not enough data for swimming's 1500 meter women's freestyle race. This is because in the 2020 Olympics, women's 1500 meter freestyle race was the first time women competed in a swimming event that was more than 800 meters. In effect, there was not previous data on the event, so the women's 1500m race cannot be compared between swimming and running.

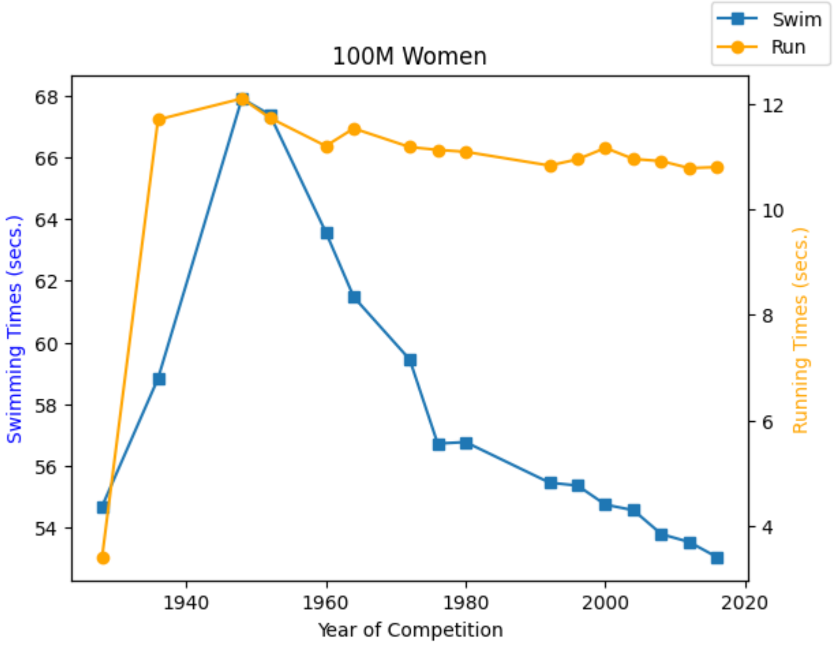

Figure 3: Olympic swimming and running results for the women’s 100 meter race

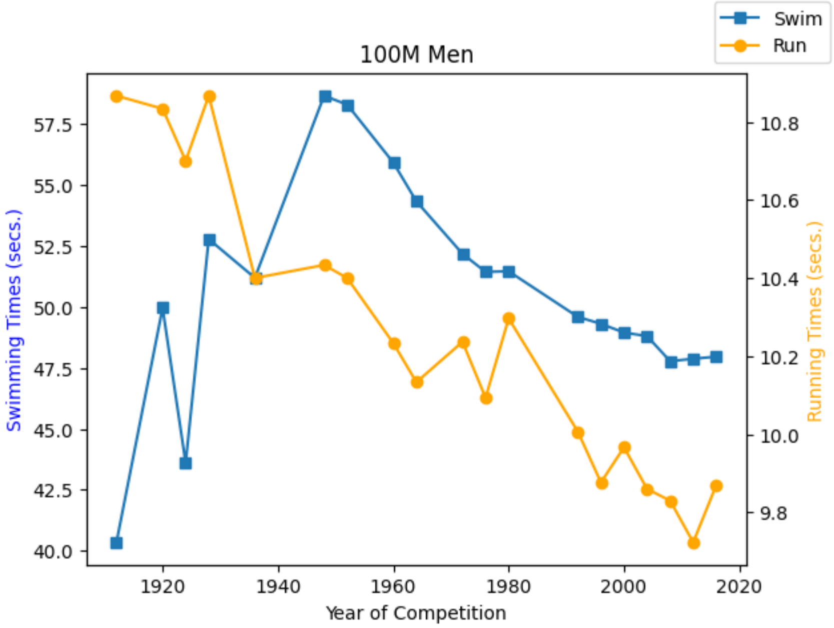

Figure 4: Olympic swimming and running results for the men’s 100 meter race

As a general statement, the running plot line seems to spike more than the swimming plot line does at first glance. However, if we take a closer to the y-axis on the right, the margin between each label is often a tenth a part from each other. This shows that even though in some years it looks like running results have significantly receded, although it is true that some years have receded over the years, it is more the fact that the difference between the surrounding data points was around a tenth, which in actuality is not a large difference.

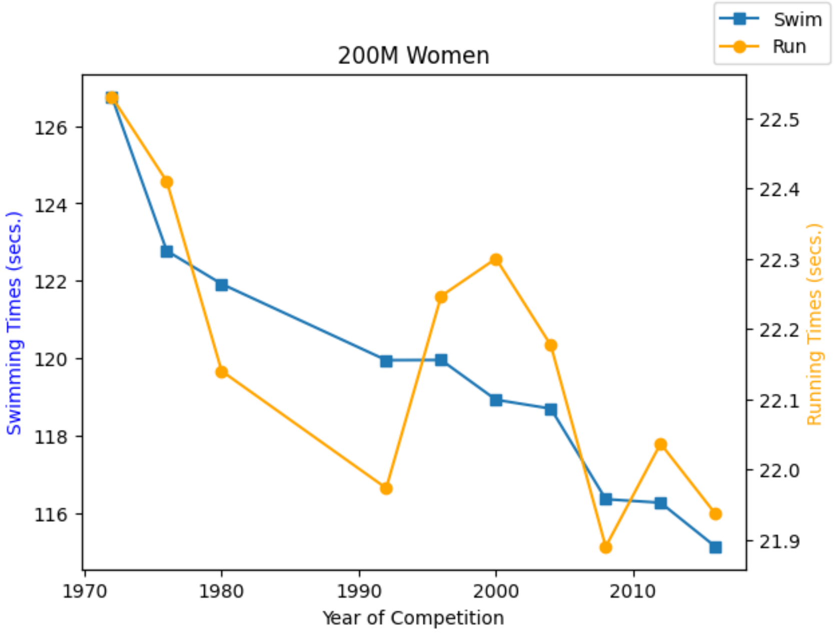

Figure 5: Olympic swimming and running results for the women’s 200 meter race

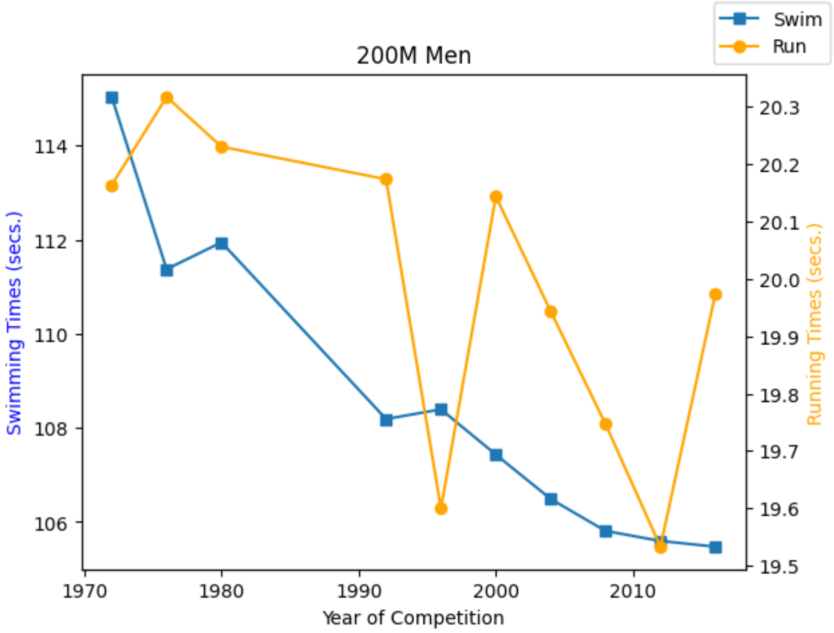

Figure 6: Olympic swimming and running results for the men’s 200 meter race

Additionally, it is interesting to note that when comparing men’s running to swimming, the men’s running trend line tends to be below the respective swimming trend line. Whereas for women, their running trend lines tend to be similar to or above their respective swimming trend line. This could be to the fact that women’s bodies are better built for swimming than men’s. For example, women have biological advantages like lower hemoglobin levels and oxygen supply, which is why women are actually better than men at marathon swimming [14]. Even though this is a small example of such and the distances are much lower than a marathon, this could still explain this pattern.

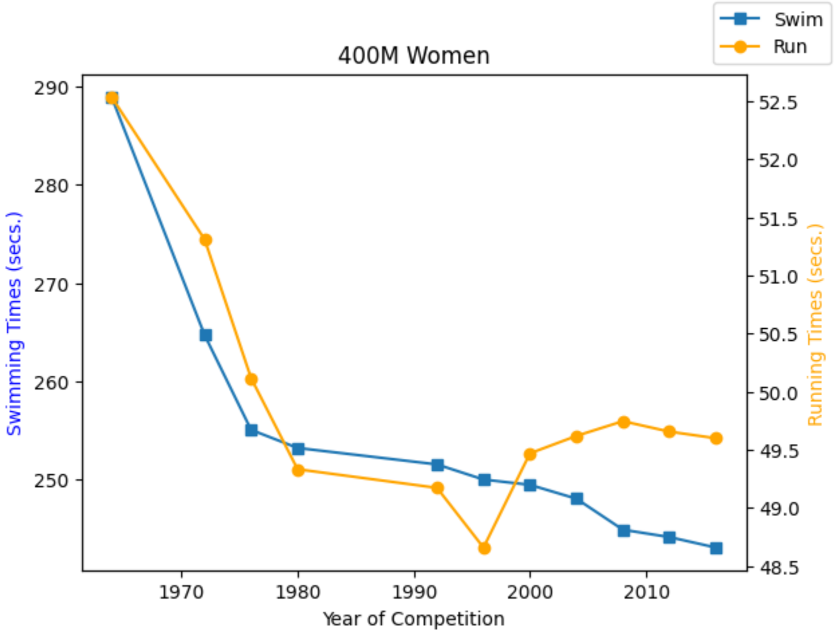

Figure 7: Olympic swimming and running results for the women's 400 meter race

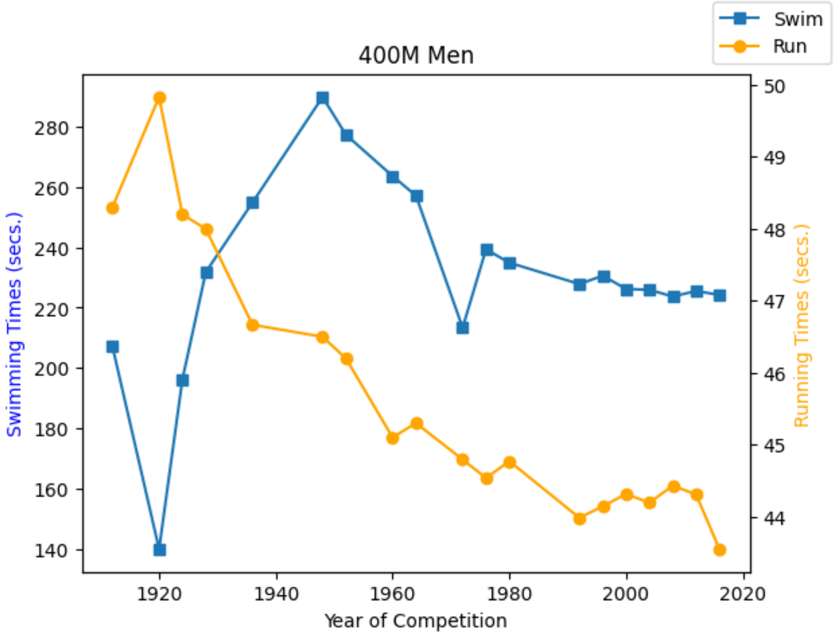

Figure 8: Olympic swimming and running results for the men's 400 meter race

Furthermore, a big spike in result times of swimming is noticeable in 1948. This is likely due to the fact that WWII had recently ended in 1945. The 1948 Olympics were there first Olympics to happen since 1936 at that time. Thus, it is likely that with world focus on the political events, recreational sports was not in the front of many people's minds.

6 Future Work

Although there was sufficient data on Olympic Games swimming results, in the future, it would be beneficial to look at additional professional swimming and running meets like the Olympic Trials, Pro Swim Series, marathons, and more. In addition to this broader data view, it would be beneficial to look at trends for different age groups. Since women and men mature at different rates, it would be interesting to view what years tend to be peak performance for different genders. These data analytics could easily be done; however, the difficulty would revolve around getting good quality and quantity of data.

Additionally, with these findings, it would be beneficial for the public to have access to such, so publishing results to GitHub would be a great result to accomplish. This could be done with expanding upon Bokeh models because the Bokeh library has many interactive features that could make a dashboard interesting and have an overall better user interface. For example, one interactive feature of Bokeh is that the viewer can choose the distance and stroke (i.e. event) of the visualization that they want to see.

References

[1] Ejeffrey M. Jones. Swimming tops track as public's favorite olympic event. 2008. [Online; accessed 6-Dec-2023].

[2] Michael phelps. *Olympics.org*. [Online; accessed 8-Dec-2023].

[3] Emma Betuel. Olympics flashbacks: How a nasa-designed swimsuit rocked the 2008 games. 2020. [Online; accessed 6-Dec-2023].

[4] Emanuele Chirico Sabrina Demarie and Christel Galvani. Prediction and analysis of tokyo olympic games swimming results: Impact of the covid-19 pandemic on swimmers' performance. *International Journal of Environmental Research and Public Health*, 2022.

[5] Patrick Pelayo and Morgan Alberty. The history of swimming research. 2011.

[6] Jose A Bragada Alan M Nevill Jorge E Morais, Tiago M Barbosa and Daniel A Marinho. Race analysis and determination of stroke frequency - stroke length combinations during the 50-m freestyle event. 2023.

[7] Joel M. Stager Chris L. Brammer and Dave A. Tanner. Beyond the "high tech" suits: Predicting 2012 olympic swim performances. 2012.

[8] David Dack. When was running invented? the fascinating history of this enduring sport. 2021.

[9] Alan M Nevill and Gregory Whyte. Are there limits to running world records? 2005.

[10] ancient olympic games. *Britannica*, 2023. [Online; accessed 6-Dec-2023].

[11] The history of olympic swimming. 2018. [Online; accessed 6-Dec-2023].

[12] Data Science Donut. Olympic swimming history (1912 to 2020). [Online; accessed 8-Dec-2023].

[13] Jay Ravaliya. Olympic track and field results. [Online; accessed 8-Dec-2023].

[14] Daniela Navarrete. Can women beat men in marathon swimming? *Swimming World Magazine*, 2020. [Online; accessed 8-Dec-2023].