DSC 540 Milestone 1

Ryan P Long

Identify Datasets

Identify 3 data sources, along with a description of each one.

Each of the following data sources contain triathlon race results from three different organizations and formats. Generally, they will contain the following variables for participants: Name, Country, Gender, Place, Swim, T1, Bike, T2, Run, Total Time, and various calculations of pace for the appropriate race segments. T2 and T2 represent the time it takes an athlete to transition from Swim to Bike, and Bike to Run. These are included in the total race time to determine the final completion time.

Data Source 1

From Ironman's boilerplate, "The IRONMAN Group is the largest operator of mass participation sports in the world and provides more than a million participants annually the benefits of endurance sports through the company's vast offerings. Since the inception of the iconic IRONMAN® brand and its first event in 1978, athletes have proven that ANYTHING IS POSSIBLE® by crossing finish lines around the world. Beginning as a single race, The IRONMAN Group has grown to become a global sensation with more than 235 events across 55+ countries. For more information, visit www.ironman.com."

Races typically consist of a 2.4 mile swim, 112 mile bike, and 26.2 mile run and are held globally during the warmer months of each hemisphere. The end of the season is loosely defined by the occurrence of the World Championship race which occurs in the early part of October each year in Kona, HI. The first data source I will be using is an **excel** file of the results from the 2019 Ironman World Championships results.

2019 Kona: 2019 Ironman World Championship: Results

Data Source 2

From the International Triathlon Union's website, "The International Triathlon Union (ITU) is the world governing body for the sport of Triathlon and its related other Multisports. It is an association founded in April 1989 in Avignon, France and existing under art. 60 and following the Swiss Civil Code. The seat of ITU is in Switzerland and its headquarters shall be decided by the Executive Board. ITU is a non-profit-making organisation and does not pursue any objective for its own gains."

Annually, the ITU hosts the World Triathlon Series (WTS) which culminates with a Grand Finale race which covers 1500m swim, 40k bike, and 10k run, typically held in either August or September. The second data source I will be using is results from the 2019 WTS final held in Lausanne, Switzerland. I will be pulling the data from the ITU's **API** for athlete results, keying in on the finale race results.

Athletes API Overview

Web based results: Results: 2019 ITU World Triathlon Grand Final Lausanne

Data Source 3

Challenge Family (CF) is a European based race organization which organizes Long (Ironman distance) and Middle (half Ironman distance) races globally, though in . The CF championship is held in Q2 annually, uses the Middle Distance format, and rotates location usually within Europe.

The 2019 championship was held on June 2, 2019 in Samorin Slovakia. The race results from both the Professional and Age Group (amatuer) races held on the same date will be pulled from a race result hosting **website**.

CHALLENGE FAMILY - THE CHAMPIONSHIP, 2019-06-02

Potential Data Source 4

From Super League Triathlon's website, "Founded in 2017, Super League Triathlon (SLT) is the pinnacle of the sport on the world's stage. We deliver game-changing race formats in a series of fast-paced events with unpredictable outcomes that culminate in the crowning of the best and most versatile male and female triathletes in the world. Over the course of seasons, the top male and female triathletes from around the world battle it out for the crown to become the best all-round triathlete."

The 2018 season is the most recently completed "season" of SLT and the final results will be pulled from the organization's **website**.

SLT Season'18 Overview » Super League Triathlon

The relationships between them, or the relationship you will make between them.

These datasets contain triathlon race results of high performing professional and amateur triathletes from championship events. I will use name and country of origin as the key data fields to create relationships between the sets.

What you believe you will have to do to the data to accomplish all 5 milestones and what your interpretation is of what the data means (you could provide a data dictionary or a summary of what the data is) – should be at least 250 words

As with most data focused projects, the time consuming piece will be the cleaning and transforming to ensure the data is consistent between all sources. These data sources are all tied together topically on triathlon, but the sources are not the same and in some cases represent opposing interests with little to gain by aligning their data outputs.

All of the data sets contain triathlon race results from different organizations and racing distances. There are some cross-over of athletes between the Ironman and CF events, however it is not likely there are athletes from either of those events found in the ITU results. ITU and SLT are primarily focused on short-course racing and switching to the long course demands of IM within the same season is possible, but typically not successful. What I'd primarily be interested in is the countries each of athletes represent and the aggregated results. I think an expanded scope project could leverage historical data to analyze the influence of olympic cycles and then the long term impact on long course racing as olympic focused athletes cycle out to longer course racing. Typically, short course athletes can transition to long course racing after a season of refocus, however the opposite is not true as long course triathletes do not successfully transition to short course. This project would be a stepping stone into long term review and analysis for confirming these statements.

I think the greatest challenge will be the extraction and formatting of the SLT event results. The website doesn't appear to be set up as a simple table. There are multiple dropdowns and tabs/pages created within the main results page. Navigating, extracting the format I need for integration and future comparison purposes appear to be the greatest challenge on the surface. I also notice the SLT results pages don't have the athlete's country of origin either, but perhaps could be applied from the ITU's results or athlete profiles.