Project 1.0 – Dataset Submission

Team members:

Matthew Young: mdy12
Matthew Harding: mmh178
Scott Johnson: saj73
Jude Moukarzel: jjm385

Data Science Question:

How do various factors lead to success in Cross Country, Indoor Track, and Outdoor Track in the NCAA?

Dataset Descriptions:

1. United States Track and Field and Cross Country Coaches Association Ranking Data

https://www.ustfccca.org/

Using web scraping via various URL paths, we collected championship data for the last 10 years for cross country, indoor track, and outdoor track. One notable variable is the change in ranking since the previous week. Here we can analyze which teams performed better in championships based on a variety of factors gathered below.

Note: to see some of the specific cross country datasets, reference the links below:
https://www.ustfccca.org/team-rankings-polls-central/polls-rankings-hub?coll=11762
https://www.ustfccca.org/team-rankings-polls-central/polls-rankings-hub?coll=11763
https://www.ustfccca.org/team-rankings-polls-central/polls-rankings-hub?coll=10446
https://www.ustfccca.org/team-rankings-polls-central/polls-rankings-hub?coll=10447

2. World Weather Online Historical Weather API

https://www.worldweatheronline.com/developer/api/docs/historical-weather-api.aspx

After cross-referencing the dates and locations of the NCAA XC Championships from https://www.ustfccca.org/meets-results/meet-history?series=3367, we used this weather API to get a full scope of the weather at the site of the championship based on date and zip Code. Combining this data with the previous data will hopefully provide some trends in which teams performed better in various types of weather.

3. Reddit API

https://www.reddit.com/r/trackandfield/
https://www.reddit.com/r/CrossCountry/

Using the Reddit API, we collected 100 posts from the subreddit "Track and Field" and 100 posts from the subreddit "Cross Country". These are trendy, or "hot" posts, which we will use to analyze and assess the public's general feeling and reaction towards the sport and what some of the most talked about factors are.

The text data was structured and formatted within a data frame, with the 2 main columns "title" and "body" containing the text content of the posts.

4. Topographical API

https://portal.opentopography.org/datasets
https://portal.opentopography.org/usgsDataset?dsid=USGS_LPC_FL_LeonCo_2018_LAS_2019

Using the topographical API we were able to download elevation data at a location of a known cross country course. The data in this repository is a sample of that data as the full dataset was too large to upload. The data using latitude and longitude to mark elevations determined by lidar technology. There are many different locations that we have access to with known cross country courses and we plan to use this data to compare different courses and performances.