Final Project Report

Clancy Andrews

12/12/2021

Cpt_S 315

Fall 2021 Final Project

2021 High School 1A Cross Country Caribou Trail League Statistics

Run Program: 1. Open path to files in terminal 2. Run python Final.py -or- python3 Final.py

Introduction

I would like to better understand the current league standings for high school cross country. Some questions I would want to answer are what the overall average time is in the league, and what is the overall average time for each grade level and team. I would also like to know, if a hypothetical meet were to take place, the rankings of the teams, i.e., who wins based off the scoring system. Another question I would like to answer is when given the ranking of all the individuals in the league, I want to find my potential standings based off the times that I ran in high school. My high school got our cross country program when I was a freshman. I excelled at the sport and found a love for running, competition, and improvement. Because of this, I often compared myself to the rest of the league, and even the state standings. Looking at the league today, I want to better understand the performance of the athletes in the region I grew up in.

Data Mining Task

The goal of the data mining task was to better understand the composition of the Caribou Trail League. Some questions I would want to answer are what the overall average time is in the league, and what is the overall average time for each grade level and team. I would also like to know, if a hypothetical meet were to take place, the rankings of the teams, i.e., who wins based off the scoring system. Another question I would like to answer is when given the ranking of all the individuals in the league, I want to find my potential standings based off the times that I ran in high school.

Methodology

The source of my data is the 2021 High School 1A Caribou Trail League rankings. The top 50 ranked spots can be viewed with an account at Athletic.net, https://www.athletic.net/CrossCountry/Division/List.aspx?DivID=61478. This is the website that contains all the Washington State High School XC and Track results. Everything is official and updated periodically. The dataset is already cleaned up, so the only preparation is to convert the columns needed into computationally usable value columns. To answer the first question, we need to calculate the average of the times for each grade level. To achieve this, we need to divide the data into different categories based off of the different grade levels. From there, we can sum up the times for each grade level and divide by the number of individuals that are present in those grade levels. The overall average for the league can be computed using all of the entry's times and the total number of entries.

The second question can be computed using the rankings and teams that are present in the dataset. I will determine the overall team rankings based off the placement of the top five runners on each team. The scoring system is as follows: Let A be the set of the top 50 runners in the league. Every team has at most 7 runners competing on varsity, but the score is determined by the top 5 runners. In the hypothetical meet, the league rankings are simulated as if they were run in the same race. From there, the top five runner's finishing position, rank, on each team is added to the team's score. The team with the lowest score wins the meet. An example of this is that team 1 have runners finish in the positions 1,3,5,7, and 9 and team 2 have runners finish in the positions 2,4,6,8,10. The team scores are team 1: $1 + 3 + 5 + 7 + 9 = 25$, and team 2: $2 + 4 + 6 + 8 + 10 = 30$. The lowest score wins, so team 1 wins the meet. This is how the scoring system works and how I will be determining the team rankings.

The final question can be determine by subtracting the league's average time by my best time. The number of standard deviations my best is away from the average can be determined by calculating the standard deviation of the league's times, and then dividing the difference of my time from the average time by the standard deviation. The output of the results to these questions will be printed to the console and an output file.

Results

My results were interesting and brought insight to what may happen in the coming years. I found that the averages for the different grade levels were: 12th grade average for 5000 meter distance: 19:49.4 11th grade average for 5000 meter distance: 19:41.3 10th grade average for 5000 meter distance: 19:32.1 9th grade average for 5000 meter distance: 20:47.0 The overall average time for 5000 meters for the league: 19:48.6 This shows that there are faster, younger individuals that will be participating next year and the year after that. When athletes return the next year, they are more likely to get PR's, personal records, and improve over the season. Based off the these calculations, I would expect these averages to decrease (times get faster) the next two years.

The results for the hypothetical meet (if the dataset where from one race) were calculated and the results are as follows: Cascade (Leavenworth) got score of 25. They placed 1 in the league by hypothetical meet and won the meet. Chelan got score of 49. They placed 2 in the league by hypothetical meet. Cashmere got score of 53. They placed 3 in the league by hypothetical meet. Quincy got score of 143. They placed 4 in the league by hypothetical meet. Omak got score of 146. They placed 5 in the league by hypothetical meet. Cascade (Leavenworth) is the predicted winner of the league, with Chelan and Cashmere have a close finish of second and third, respectively. Quincy and Omak where close, but Quincy prevailed and Omak placed last in the league. It is not uncommon for these standings to sway as new season best times for each individual get updated into the database. As of the point when I extracted the data, these where those standings.

The final result was comparing my best time with the league standings. My best time in high school was 16:52.7, which I achieved at the 1A State Cross Country Meet with a placement of 31st out of 220 competitors. The following are the results of my calculations: My best time for the 5000 meters was 16:52.7. That time is 175.9 second(s) faster than the league average time. My time is 7.71 standard deviation(s) from the average time of the league. If I were back in high school at the same fitness level I was, I would be able to easily compete in the league. However, when I was competing in this league, I was racing against two younger runners who ended up getting multiple state championship titles a piece. It is worth noting that these numbers will be slower due to the current COVID-19, but seeing the current trend in average times, I am confident that there will be faster runners that will dominate the league again.