Technical Report

Abbey Kays, Braedon Fiume, Camden Heafitz, Emma Strawbridge

2023-05-08

Abstract

This report considers data from the Track and Field Results Reporting System database from 2010-2022 and attempts to determine whether 2022 DIII athletes, the fastest or average national qualifiers, would have been able to qualify for DI competition in the 100m, 800m, high jump, and shotput events. This study analyzes regional trends of the top 20 qualifiers for DIII schools, comparing six years between 2010-2015 and six years between 2016-2022 counting the total top 20 qualifying athletes per state in the 100m, 800m, high jump, and shotput. Finally, this study aims to predict NESCAC results for the 2023 racing season based on 2010-2022 results for the men's and women's 100m, 800m, high jump, and shotput.

Introduction

The age-old question of "Could I have gone DI?" plagues youth minds to this day. With the rise of competitive college athletics, this question is asked by more and more high school students, washed-up college graduates, and current DIII athletes tired of losing. A second but still curious question is, "Why is Wisconsin so fast?". These heroic statisticians seek to answer this inquiry. And, as a bonus, they tried to predict NESCAC results for this season.

For the question of "Could I have gone DI," we determined that there was necessarily a definition of what "going DI" might have been. In order to make this as robust as possible, we defined this as being skilled enough in a given event to qualify in the top 20 places for DI nationals. This way, one would be able to defend being good enough to be on a DI team because one would actually qualify for nationals. Simply put, the claims would be substantial. There are plenty of DI athletes that do not qualify top 20 times but go to the Olympics, but the evaluation is not on the Olympic level. The goal is to find out whether athletes would be successful in a DI environment. For regional trends, given Amherst's NCAA division, we only investigated DIII trends as that would be the information relevant to our team. The top 20 qualifiers for national championships were once again investigated, and the total number of qualifiers in the top 20 would determine the state's strength as a team. Finally, to determine NESCAC results, getting even more specific to Amherst's regional division within DIII sports, we analyzed the years from 2010-2022, excluding years where the event was not held. For these results, only the winning time or mark was analyzed.

Track and field is comprised of 21 different events. There was not enough time or ability to analyze data from every event, so significant cuts were made. We decided that a short sprint, a middle-distance run, a jumping event, and a throwing event would be used. The final selection of events was the 100m, 800m, high jump, and shotput for men's and women's teams. All results were derived from these four events. The year 2020 was omitted from all analyses because the season was cut short due to COVID-19, so qualifiers would have had significantly less time to compete and improve, and the NESCAC competition was not held.

Methods and Data

The data was obtained from the TFRRS database, a publicly available and school-updated track and field results database. It contains information from 2009 through the current year in every NCAA-sanctioned track and field event. For this study's purposes, times and marks from 2010 through 2022, excluding 2020, were used from sections indicating qualifying times for NCAA national championships and NESCAC results. The data is organized very well on this site and easy to scrape from the DI and DIII qualifying times list, but the NESCAC collections are a little bit of a nightmare. The index of the tables is different between years, so scraping this data was difficult. There was also a set of data manually created from school data to match which state the school was in. This information was not available on the site for the schools, so it had to be handmade.

In terms of the scraping, dates, times, and marks were modified to be datetime or numerical objects, respectively, so everything would be usable to perform analysis with. The state data was combined with DIII data, DI/DIII data with itself, and NESCAC on its own. For the analysis of the data, based on what was discussed in the introduction, comparing the mean and "fastest" (either max for distance or height and min for time) times to each other proved to be the best way to find the desired answer. The maps needed some simple spatial analysis, and because there were only two variables present and usable for the NESCAC data, time and state, a simple linear regression was appropriate.

Results

The first set of results is those that answer the question of whether or not DIII athletes could have "gone DI,"already defined as being skilled enough to qualify for DI nationals with a top 20 time. The majority response is that no, DIII athletes could not compete at the DI level, but there are a few athletes where this is not the case. The fastest men's 100m DIII athlete in 2022 could have gone DI as soon as 2021 and would have been an average top 20 qualifier in 2014, as seen by the place in which the dashed line intersects the DI trend line, the average of the DI times (Figure 1). The fastest women's 800m DIII athlete in 2022 could have gone d1 at any point during her career and would have been an average top 20 qualifier in 2014 (Figure 4), similar to the men's 100m condition. Finally, the furthest men's DIII shot put mark in 2022 would have just barely been a top 20 DI qualifier in 2011 or 2010 (Figure 7). All other conditions on events – average 2022 times or marks and fastest 2022 times or marks – would not have, in any year between 2010 and 2022, been able to "go DI" based on the stated definitions (Figures 2, 3, 5, 6, and 8).

Figure 1

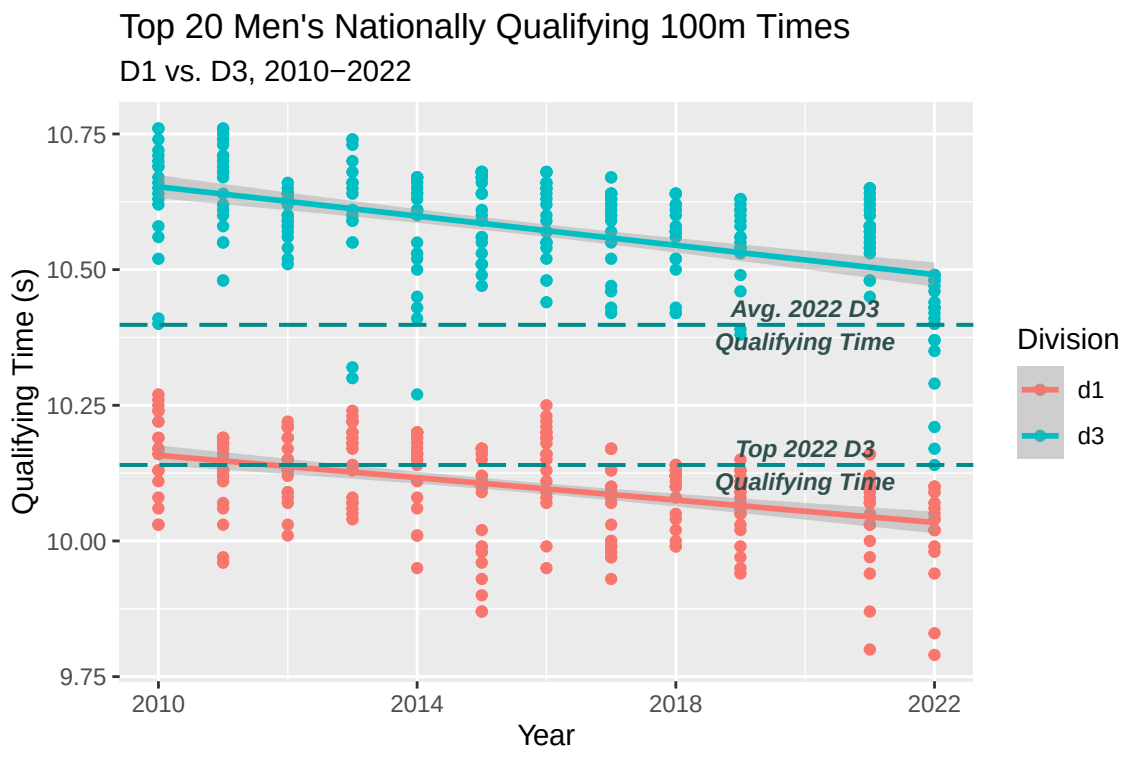


Figure 2

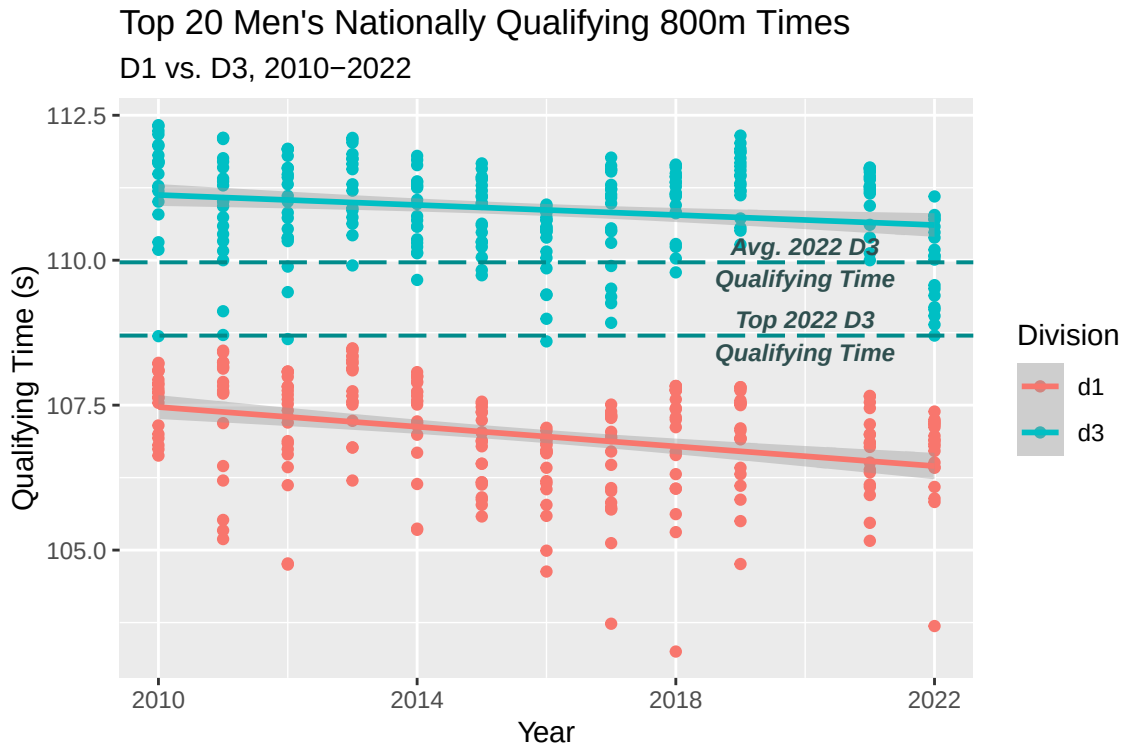


Figure 3

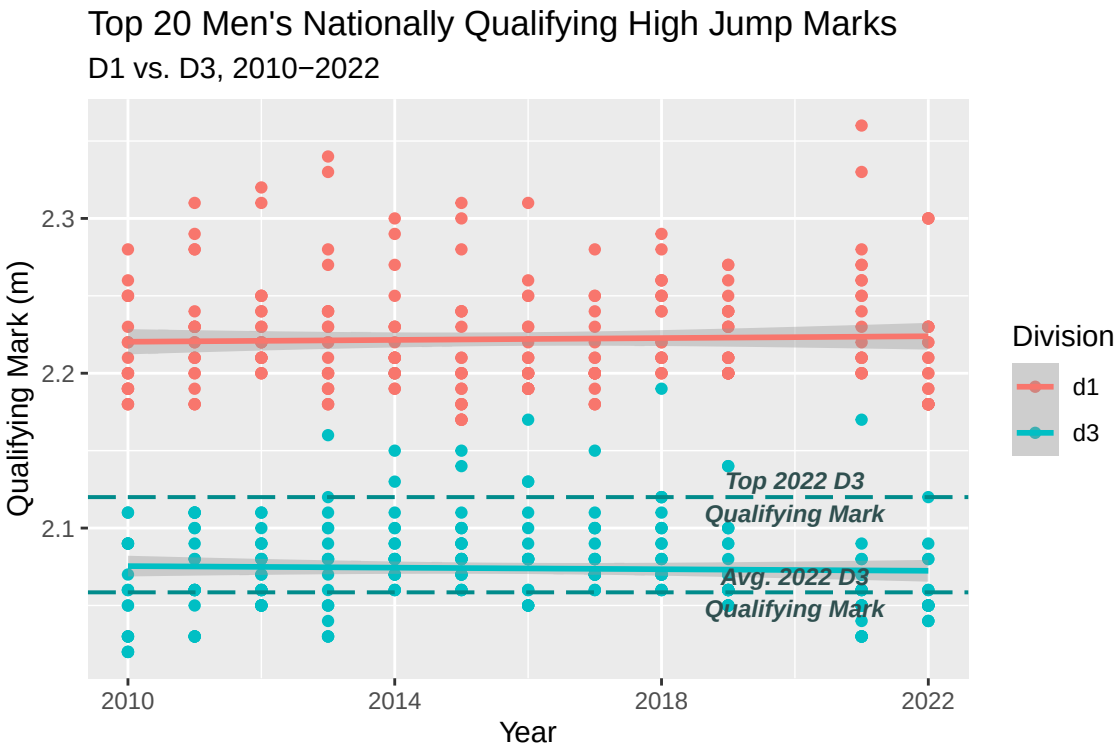


Figure 4

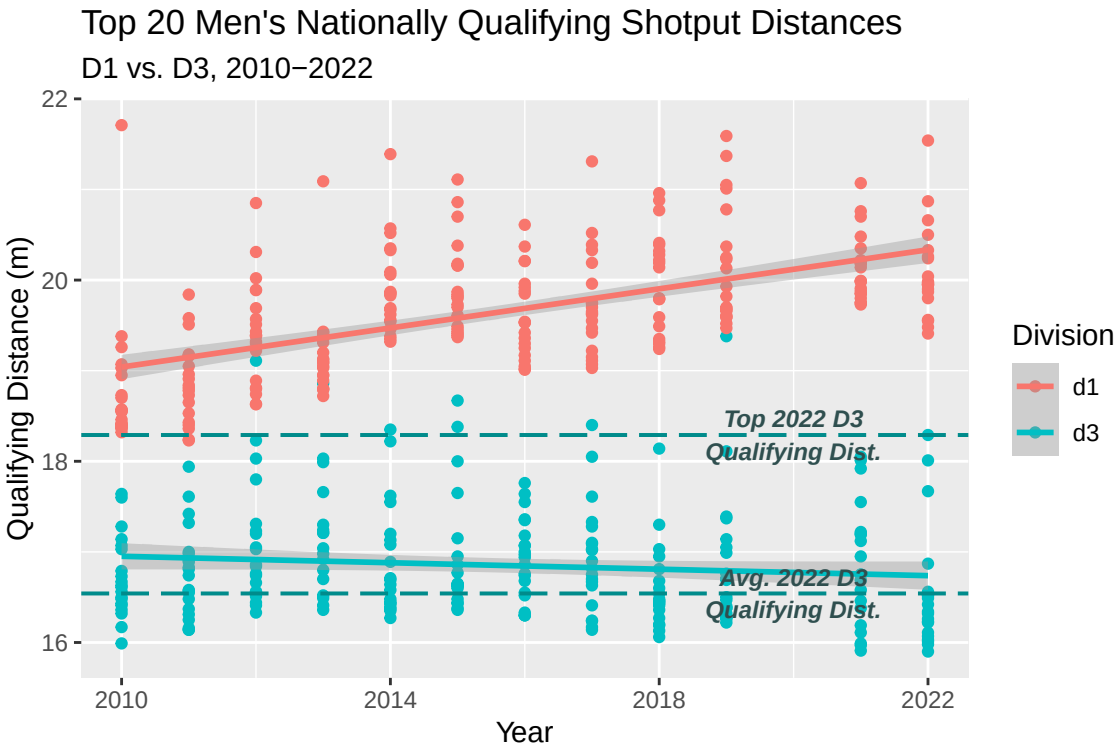


Figure 5

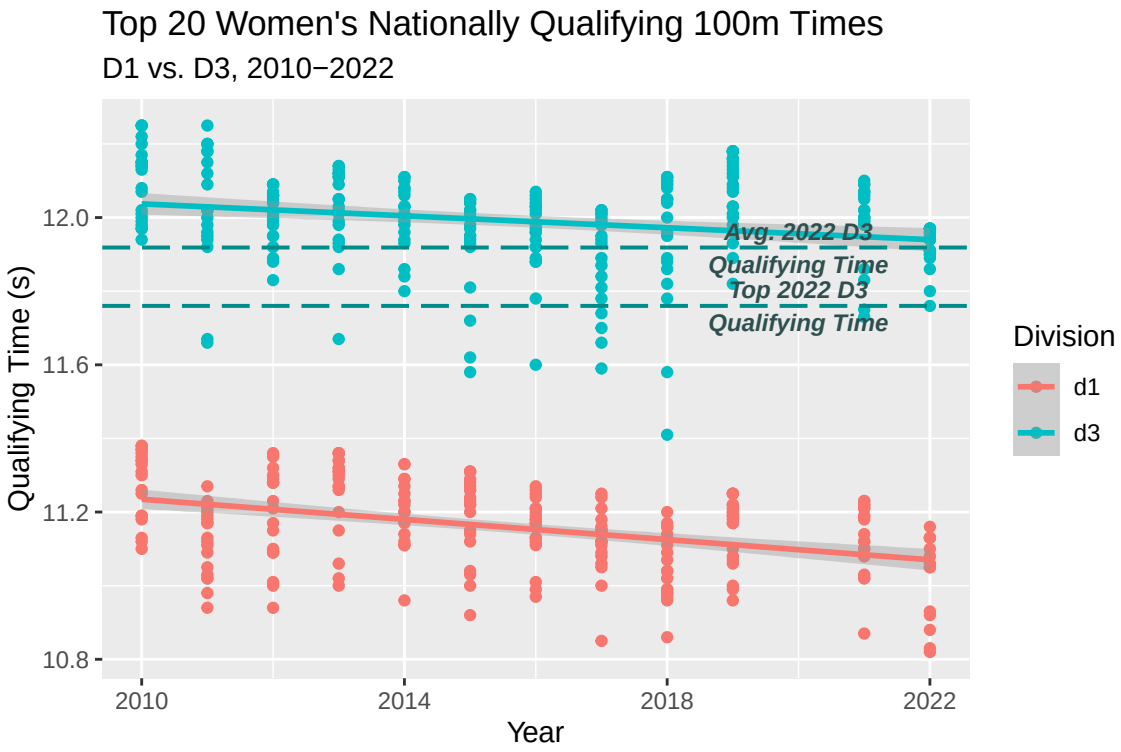


Figure 6

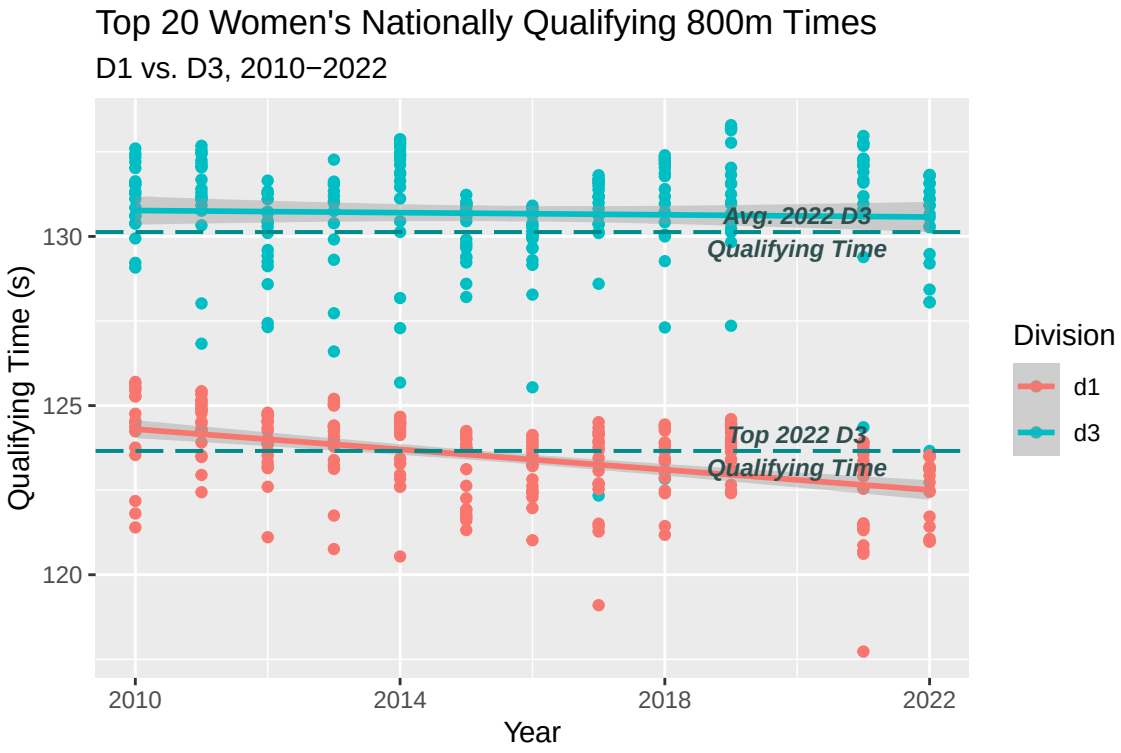


Figure 7

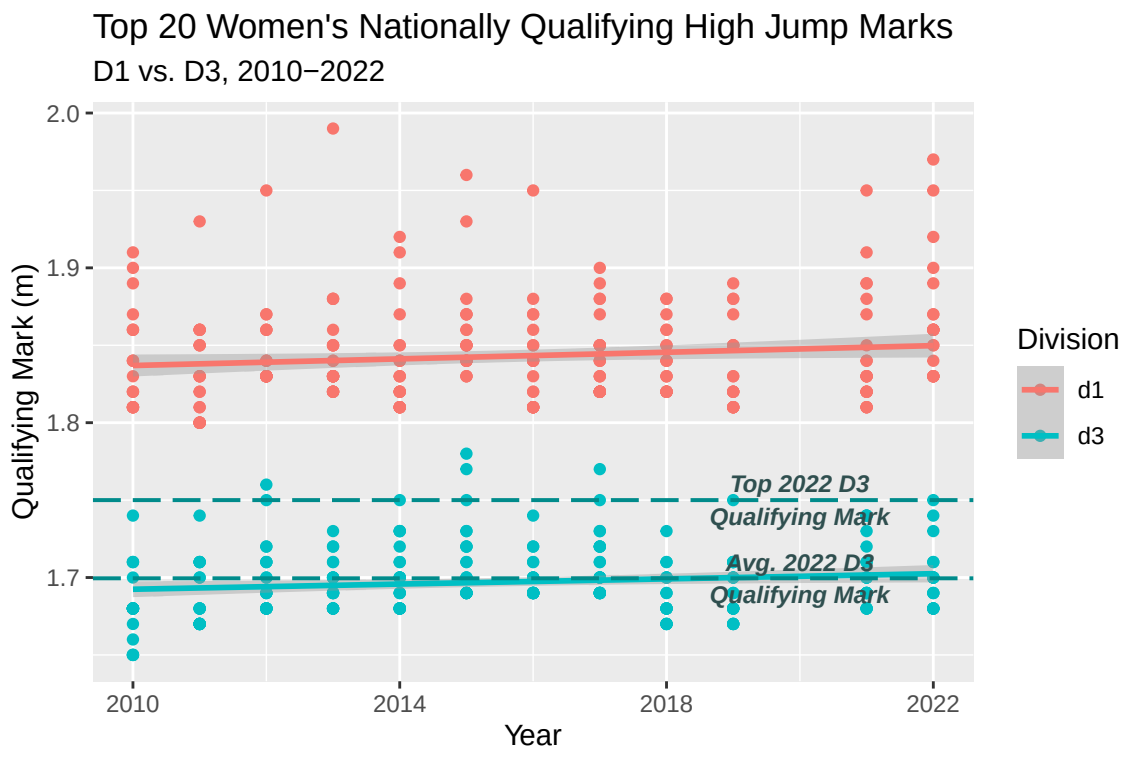
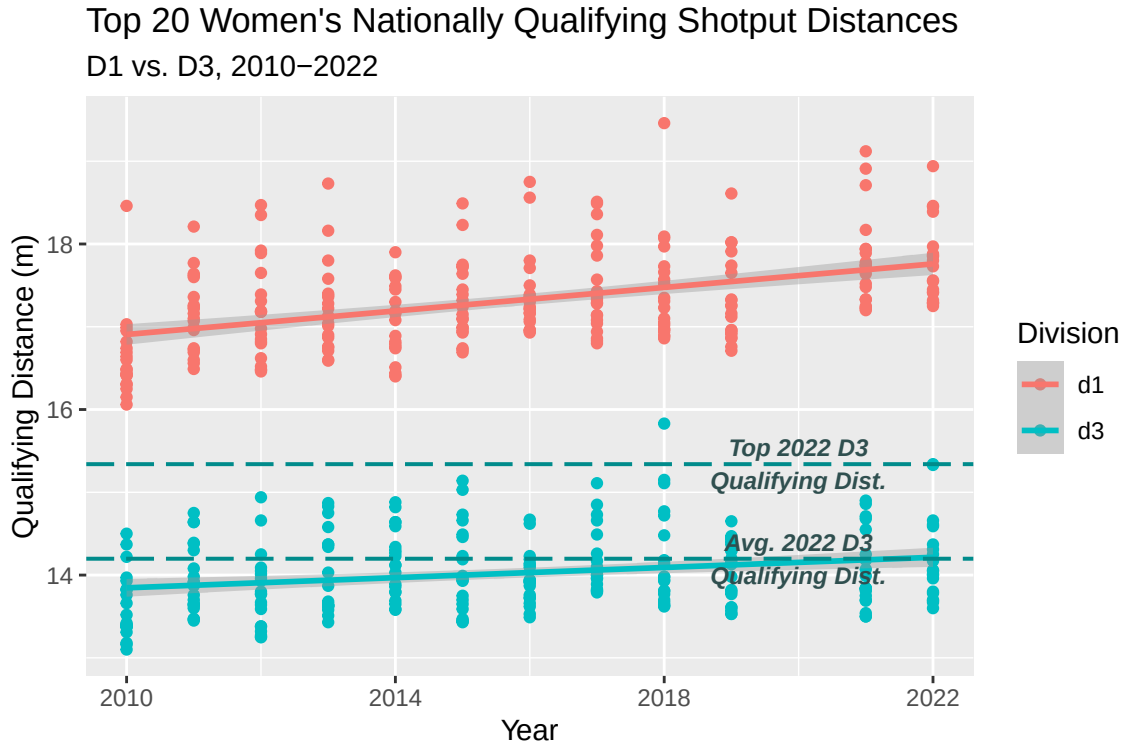


Figure 8



The second set of results concerns regional trends of DIII teams. Women's teams across all years had Wisconsin as the dominant state, meaning that the state of Wisconsin (being the combination of student-athletes attending DIII schools in Wisconsin competing in track and field) had the highest number of top 20 nationally qualifying DIII athletes across the four events that were investigated. For the first six years, midwestern states, especially Iowa, were dominant and had larger numbers than other represented states, and in the second six years, California emerged as having a high count of qualifiers, as well as Ohio (Figure 9, Figure 10). The Northeast was similarly dominant across both the first and second six years. Men's teams also had Wisconsin as the clearly dominant state across all years. In the second six years investigated, more states gained athletes entering qualifying for DIII nationals, seen by the increased number of states on the second map and especially those in yellow. In the second six years also, the Midwest collection of Minnesota, Iowa, and Illinois becomes more concentrated. Across both the first and second set of six years, the Northeast states of Ohio, Pennsylvania, New York, and Massachusetts remain strong (Figure 11, Figure 12).

Figure 9, Figure 10

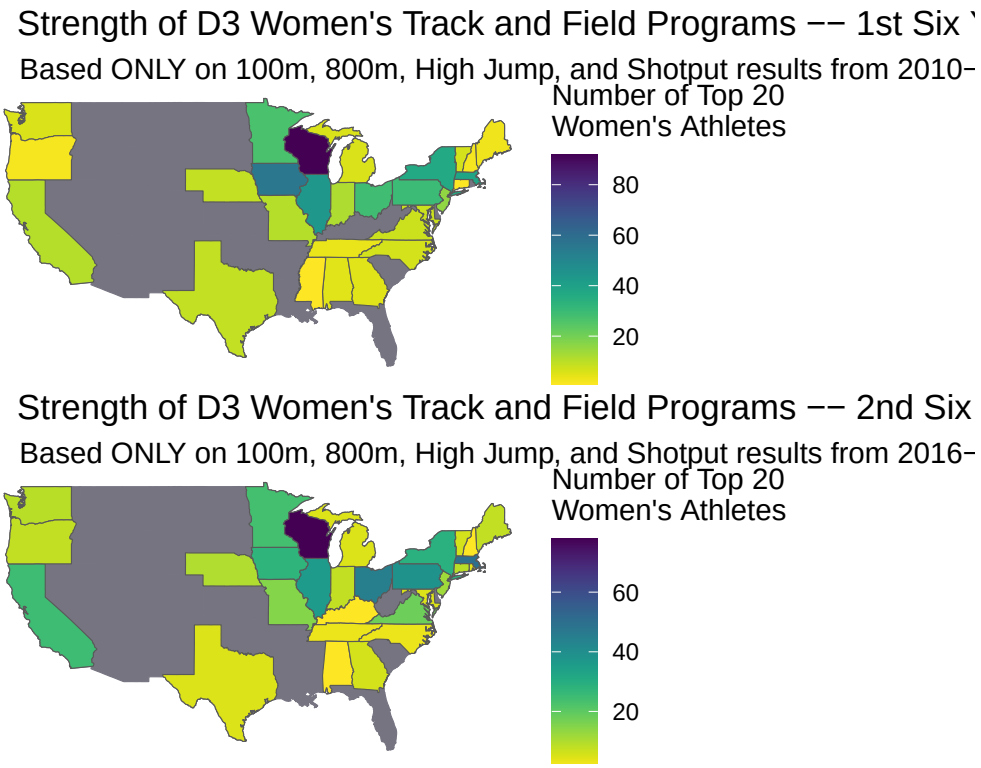
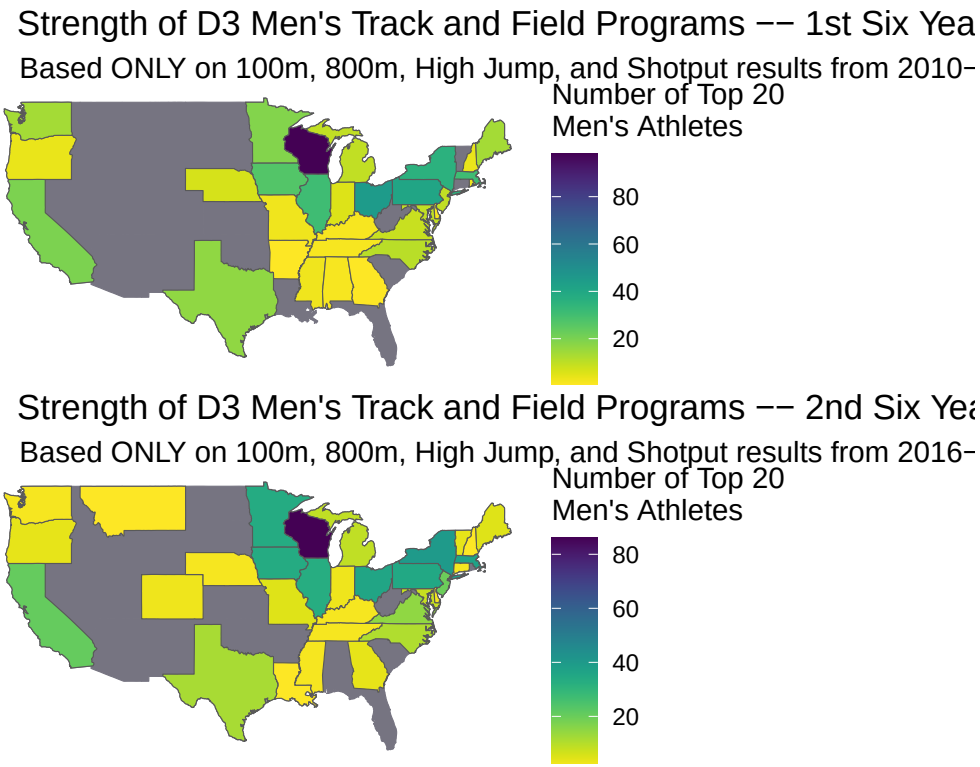


Figure 11, Figure 12



Finally, we consider the NESCAC results – those we predicted, and the actual 2023 results. All but one of the predicted times or marks were better, i.e., either faster or higher or further than the actual NESCAC champion's time or mark. One prediction, the men's shot put distance, was underestimated by the model. Figures 13-20 indicate the linear model created from time or mark and year with a green dashed line and two labeled points indicating the predicted and actual times or marks in 2023. The closest predictions were in the men's and women's high jump, with the men's prediction being 2.04 meters and the actual winning height for 2023 being 2.00 meters (Figure 17), and the women's prediction being 1.71 meters, and the actual winning height for 2023 being 1.70 meters (Figure 18).

Figure 13

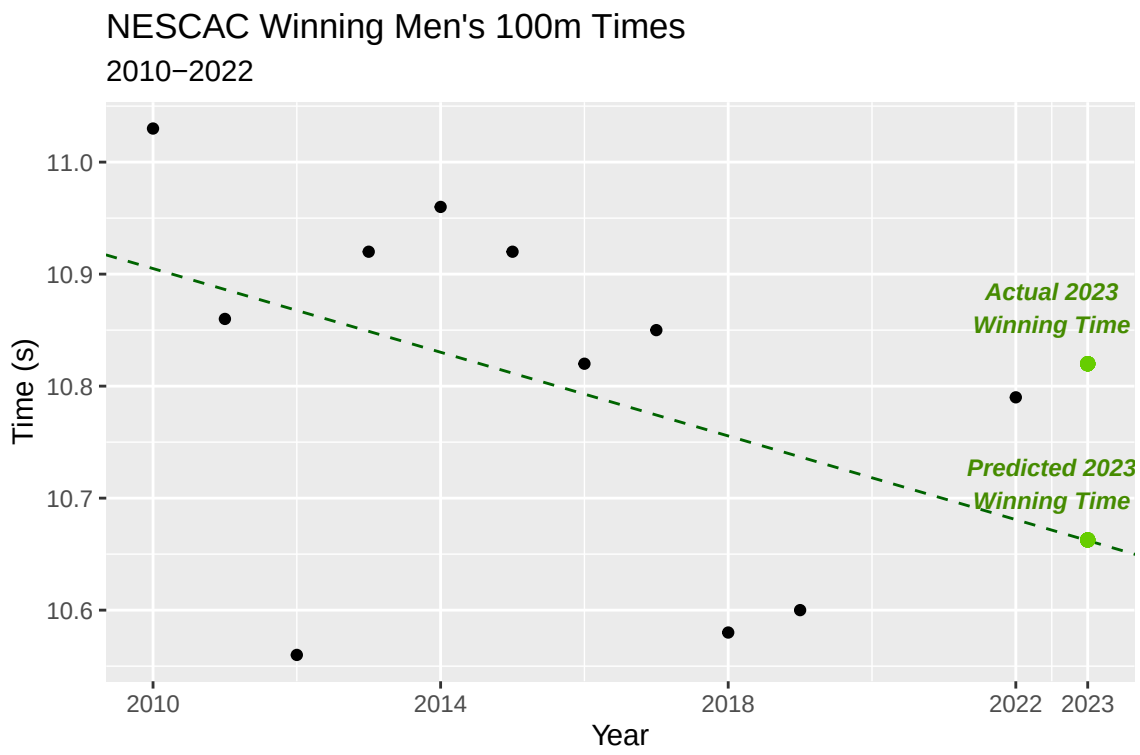


Figure 14

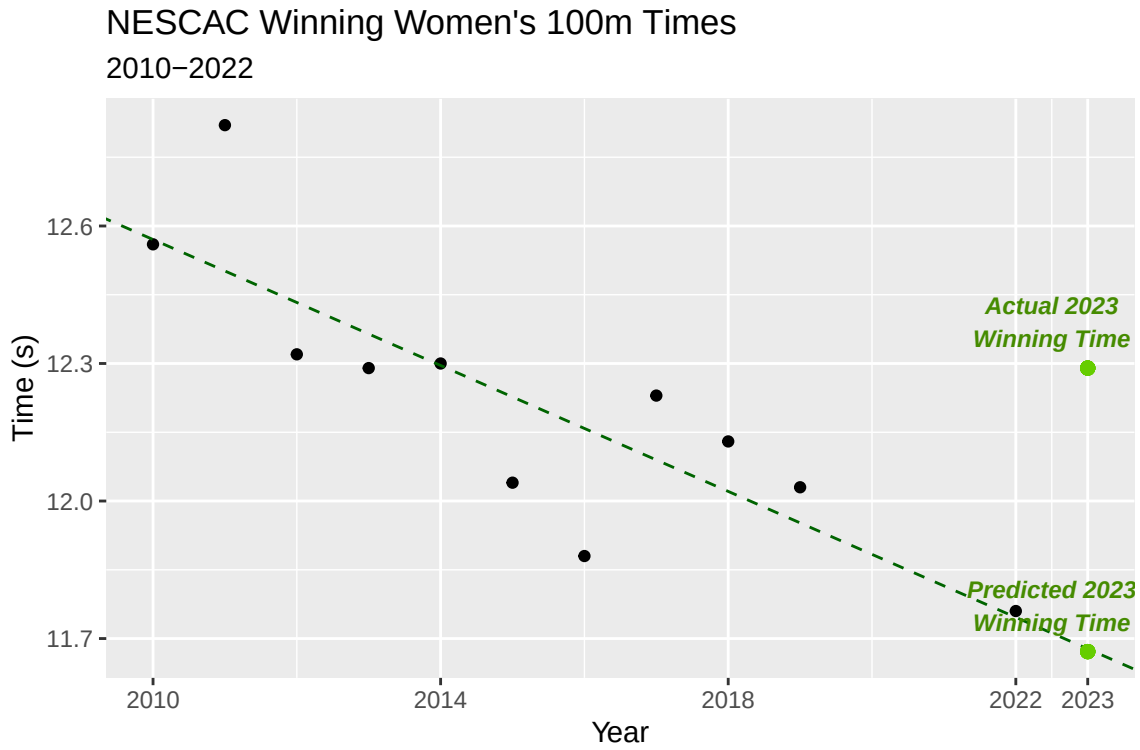


Figure 15

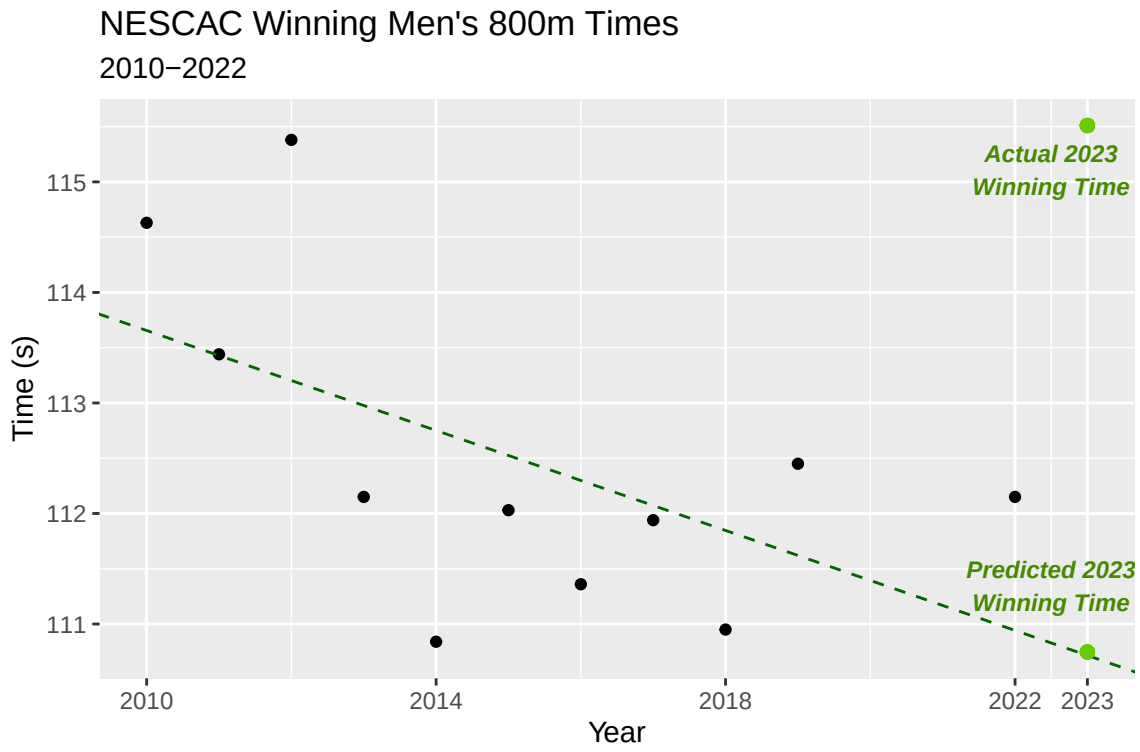


Figure 16

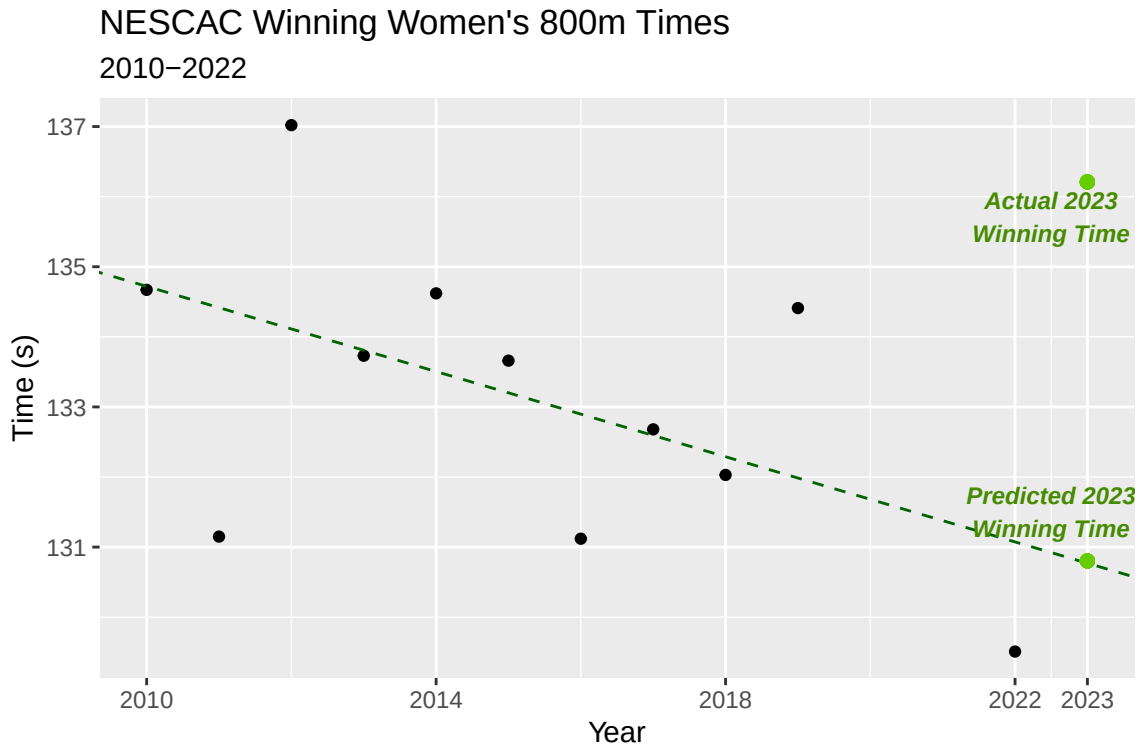


Figure 17

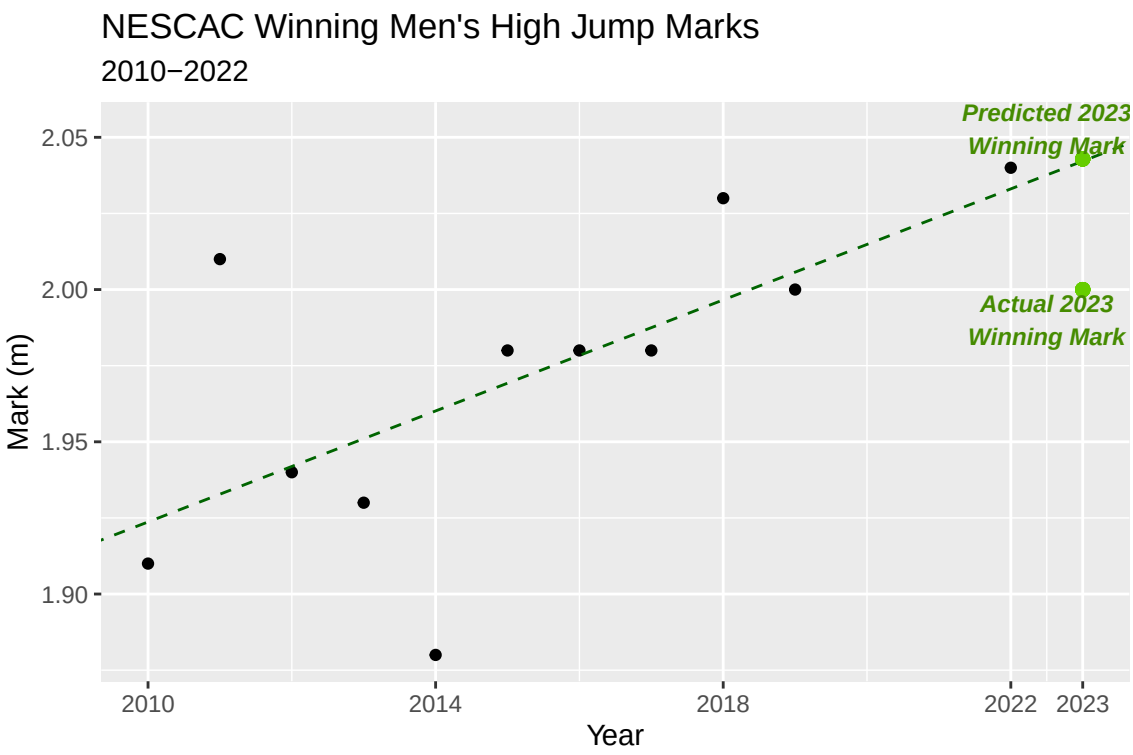


Figure 18

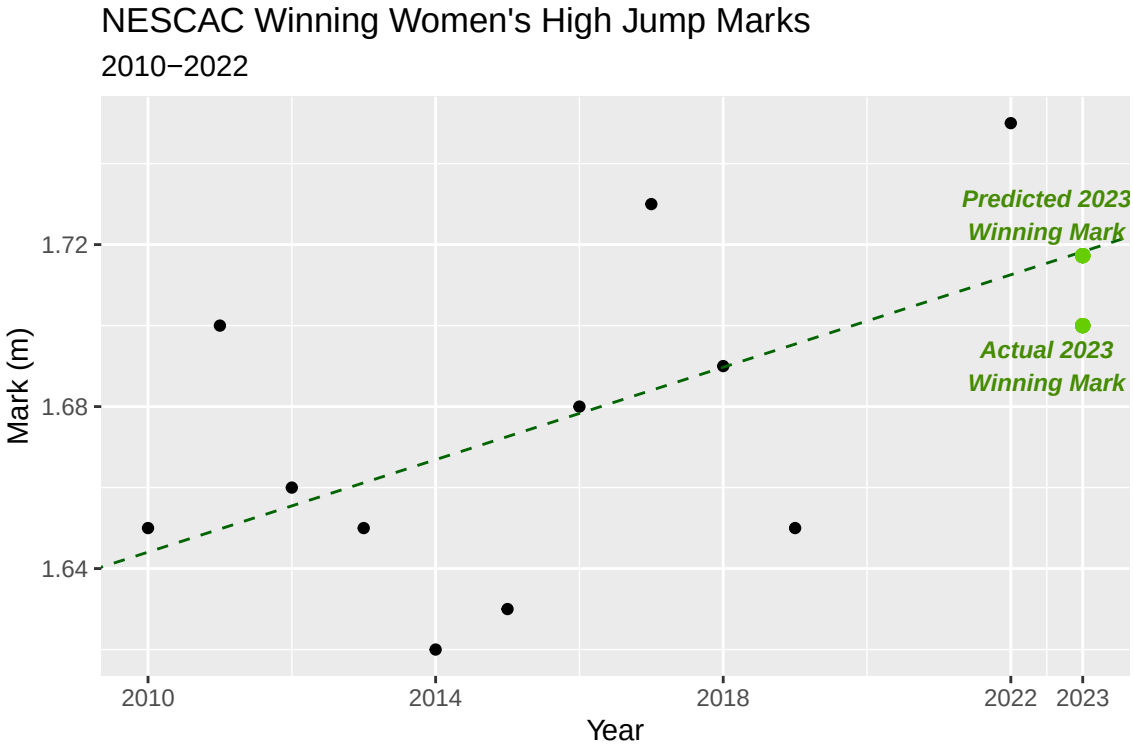


Figure 19

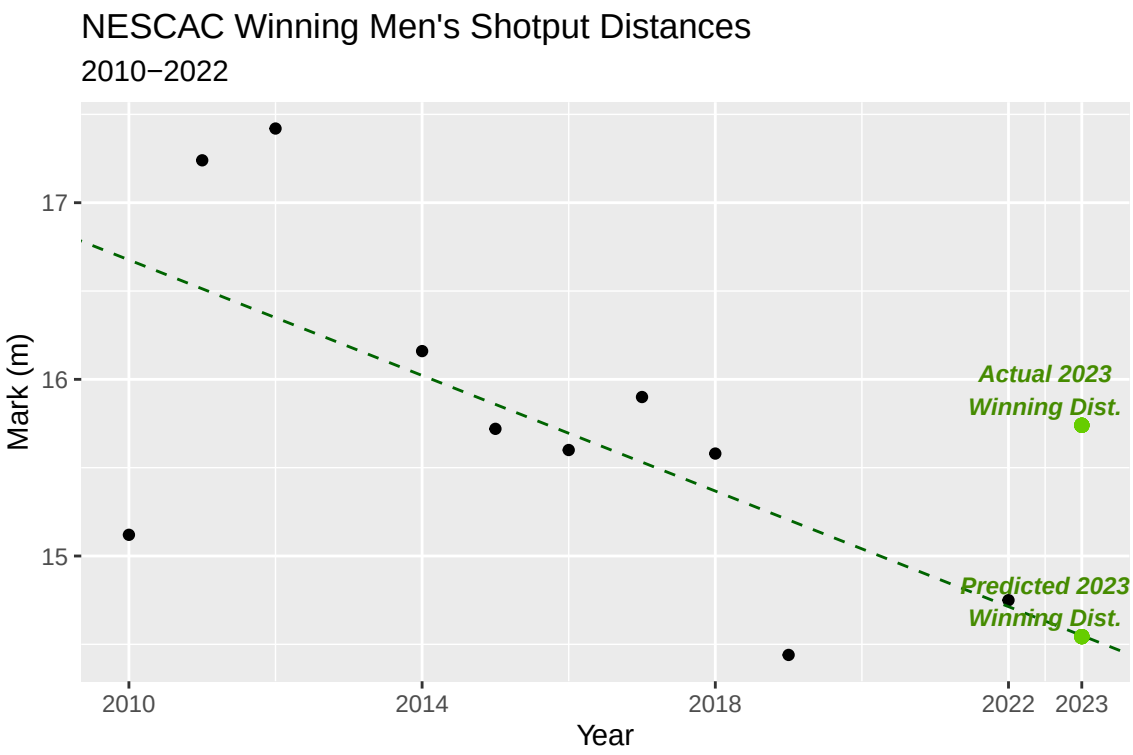
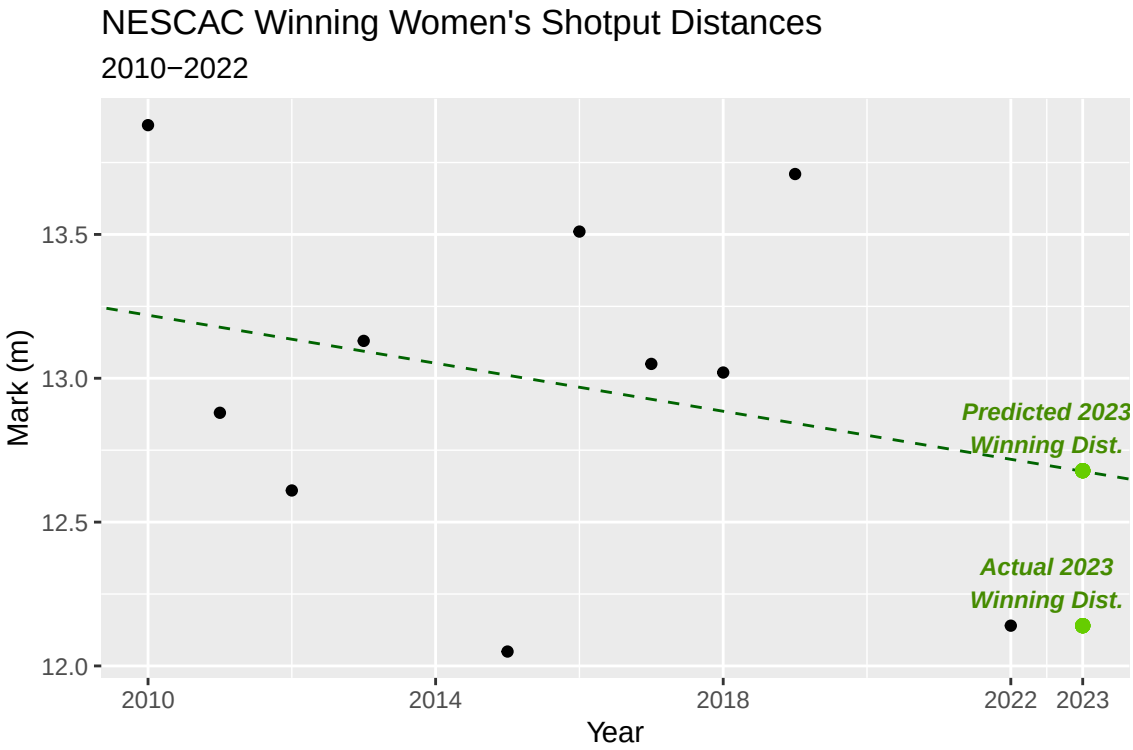


Figure 20



Diagnostics and Discussion

The first concern with our model that was necessary to consider was with the NESCAC results, namely, why they were so inaccurate and why the model's prediction so often was significantly better, i.e., faster, higher, or further, than the actual times. One of the reasons this might be the case is that the weather, particularly the wind, for 2023's NESCAC competitions was terrible. The NESCAC is split into two days of competition, and the first day, which included the most accurate prediction of the high jump, had the best weather. The second day had terrible winds and rain, and it is very possible that this skewed the times or distances dramatically. Unfortunately, NESCAC entries do not indicate the wind speed at the time of the event, while data for nationally qualifying times do, so there was no data to create a model including wind speed or some other type of weather metric.

Another concern with our NESCAC models was the diagnostics, or more particularly, the lack of diagnostics. Over half of these linear models did not have significant predictors and did not meet any standard statistical conditions, so they would not be recommended to be used in practice. Predicting times for 2023 would also be extrapolating from data, which is also not a statistically sound method for doing anything. So this NESCAC prediction process was mostly an exploratory investigation into what might or might not work well for attempting to predict NESCAC times. The actual statistical methods that are being used are questionable at best but interesting nonetheless. Our other methods of data analysis involved comparing means for our DI and DIII comparisons and comparing counts for the regional DIII maps, which do not require diagnostic tests.

Conclusion

With the visualizations created, all three questions are answerable to some degree. For the question of "Could I go DI," it depends on the event, but one could know their time and find a point of reference within these visualizations and be able to answer the question and defend it. The answer would not be sure because there are plenty of DI athletes who do not qualify for nationals and also plenty of DI schools who have no athletes skilled enough to qualify for nationals and tend not to be as skilled as most other schools. However, when this question was imagined, it felt necessary to give a robust, defendable answer to one's peers. For the regional trends, the question was simple and the answer was simple: Wisconsin DIII schools are really good, and continue to be good. Finally, for the NESCACS, there is a lot left to discover and add. The predictions were fairly close in good weather but were worse as the weather worsened, which indicates that weather should undoubtedly be taken into account when making predictions about time. Finally, there are major limitations to this project and the work put into this report. Only twelve years' worth of data and losing all of 2020 reduces the number of data points available to predict upon, especially for NESCAC predictions which sometimes had only eight points with which to create a model. Similarly, these grand statements about the strength of a team in the regional analysis are only based on four events. There are many schools known for their relays (like Amherst, going undefeated in the last three years in the 4x800, of which this year's relay Braedon was a member) or their jumps or some of the longer-middle-distance events that were not included. In order to get a fully accurate picture of collegiate team strength, all 21 events should be included. All of this being said, the exploration of this data was fascinating, entertaining, and immediately relevant given the proximity of the NESCAC events at the end of April and two of the statisticians being track and field athletes themselves. There is much to add to this study in the coming years and it will be fascinating to see how this type of analysis can grow and change.