Gymnastic Outcomes Final Report

Kirsten Dawes, Daphne Magsby, and Naveli Shah

***Abstract*— When a gymnast wants to progress into the NCAA or the Olympics, she must pass through ten levels to make it to the elite. However identifying who will succeed, or which gyms promote success is not currently possible. We want to discover factors that indicate the likelihood that a gymnast will be in a pool of top candidates based on gymnastic competition scores throughout the levels.**

I. INTRODUCTION

testtest Many competitive sports have similar characteristics to gymnastics as an athlete builds a career. One important thing for all competitive athletes to know is which location is the best for advancement, as well as which places have lower injury rates. If the NCAA or the Olympic organization can help collect data on gyms or other training locations about injuries and competitive progress, it will build the sports as well as further athletes longevity. Having data to rank training facilities will help guide the growth of new facilities.

II. DATA

USA Gym

1) Data

: To gather information necessary for the analysis the website usagym.org was used to gather information about each gymnasts score in events through the years 2008-2017. The event name and date, gymnast name, vault, bar, beam, floor, and overall scores were stored in the database.

2) Methods

: Python was used to gather the data. Packages selenium, pandas, and tabula-py were used to navigate the website and convert the pdf results into csv files for easy adding to a collection in the database.

3) Issues

: Because the usagym.org website contained data that would only become available in the sites html when it was clicked, a simpler approach to scraping the site using python package BeautifulSoup was not able to be used. Due to this, package selenium was used instead. Using selenium caused several setbacks. The first being limited examples of how to get the needed data using selenium, even with using the manual. Eventually the data was gathered, but only a small subset due to the time constraint.

*This work was not supported by any organization

[1]H. Kwakernaak is with Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, 7500 AE Enschede, The Netherlands h.kwakernaak at papercept.net

[2]P. Misra is with the Department of Electrical Engineering, Wright State University, Dayton, OH 45435, USA p.misra at ieee.org

III. DATA

Meet Score Online

1) Data

: With this sight we want to gather multiple type of data. With the focus on regions and top 100 gymnasts. This data was parsed from the website http://www.meetscoresonline.com. There were 4 types of data that were parsed gymnast , gyms, regions and competitions.

A gymnast consisted of a name , unique id , gyms , years , top100 ranks and top 100 scores (could appear multiple times).

A Gym consisted of a name, a region number, students, state , and address.

A Region consisted of Region number, and states.

A Competition consisted of competition name , date , gymnast id, bar score, vault score, beam score, floor score, all-round score , and overall rank.

2) Methods

: To gather the data I used python3 with a couple different libraries beautiful soup to parse the data and response to get the data from different parts of the sight. Once able to get the data, we place it into the mongo database for each type of data. Next using python to created need csv to do the calculations on allowed for formatting correctly and quickly. Once having the csv, I used R for the analysis as I like R studio's interface.

3) Issues

: Parsing was a large issue as there is not consistency between gym names. Example is Team Dynamics from Arizona is also inputed as AZ Dynamics or Arizona Dynamics. Parsing was also an issue as we had to know what we wanted to find. Since the formatting of the html was not nice and there was not json to parse from. Regions was a hard piece to find as there was not link saying this was region. When accessing the gym information there was a notice that not all gyms were include and not all gyms had gymnast. There is also the issue that gyms only include that years gymnast not a yearly total of there gymnast.

IV. DATA

USA Gym

1) Data

: The third set of data came from the website: http://www.mymeetscores.com ¡http://www.mymeetscores.com/¿. This website contained the Top100 people with rankings ranging from levels 3-10

and XB-XD, from the years 2009 2017.
There was a total of two types of data that were parsed: 1.Each individuals information including all the gyms the person attended, and the individuals recorded scores from competitions. 2.Gym count meaning the number of people in a specific gym that are/were in the Top100.

2) Methods

: To scrap and parse the wanted data from the website, I used Python3 and two pre-defined libraries in Python3. Beautiful Soup was used to parse the data, and Response to get the data from defiant parts of the sites. After getting the parsed data, I stored the information in a dictionary format and then exported it the data under the names MMS TOP100PEEP and MMS PEEPCOMPETITION in FDAC17FP using Mongo. I then extracted different sets of data needed for the analysis from the database, and then exported this information into a CSV file using R. Finally, I created bar graphs for the data in the CSV files using Excel.

3) Issues

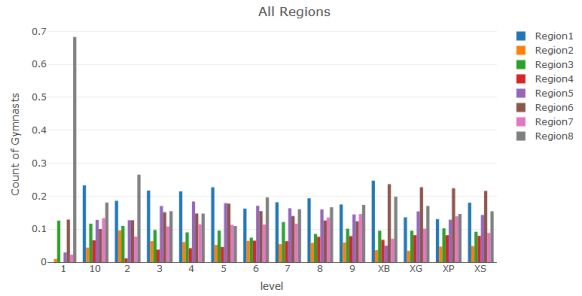
: This was my first time attempting to parse data with Beautiful Soup. This meant that understanding how it worked took a little time. The website used for this part of data had a lot of information and to go through every page that I was using to get the data took at least 2 hours each. This made debugging very time consuming. Plus, tables on this website were not formatted correctly using html. Therefore, getting specific data from the pages took a lot of thought and debugging.

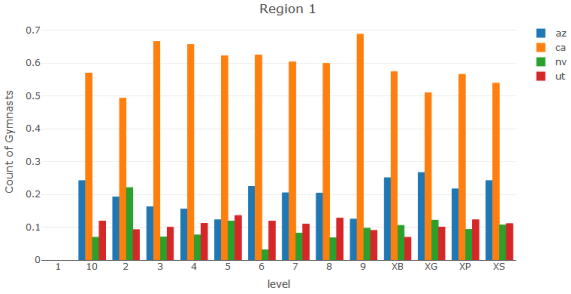
V. RESULTS I

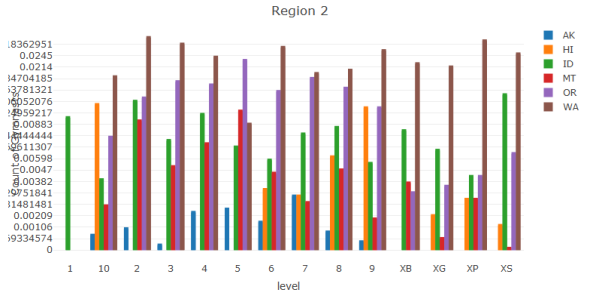
The questions we wanted to ask were :
Do region or state locations impact a gymnast success?

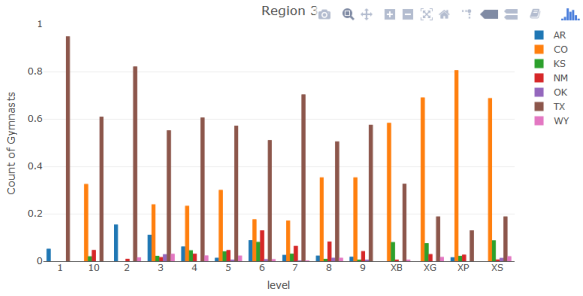
Region or state impact

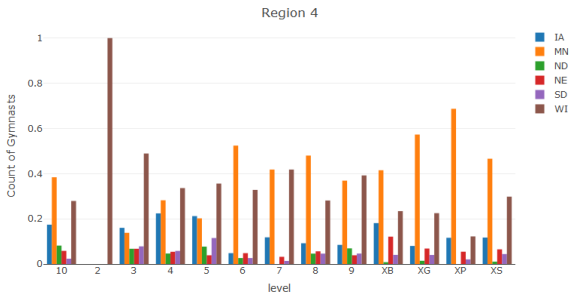
percentage of Gymnast in each Regions

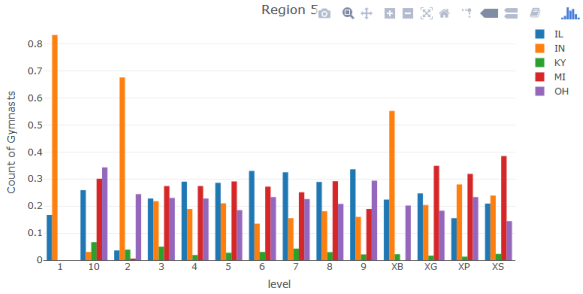


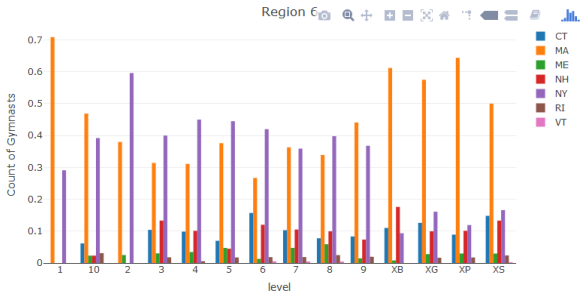


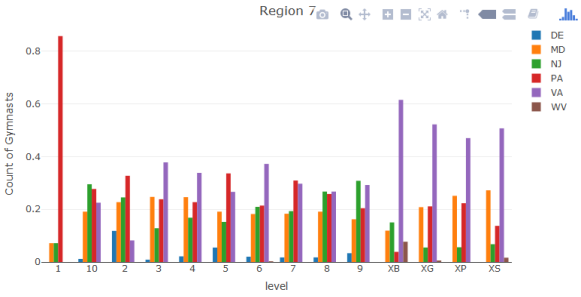


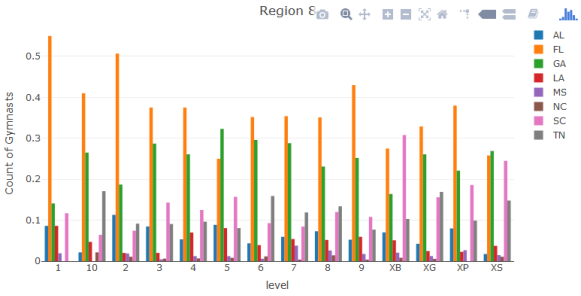


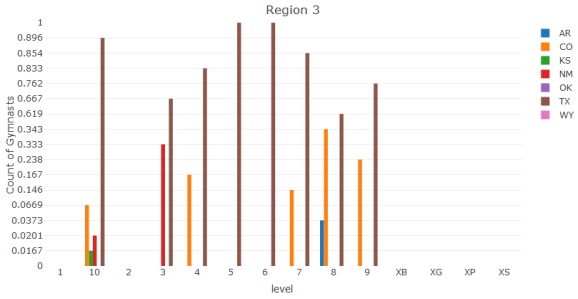


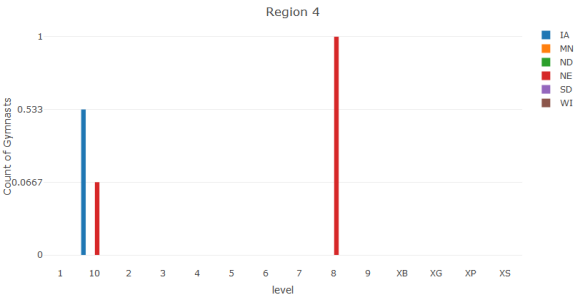


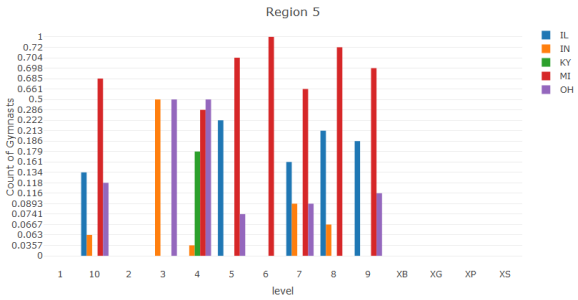




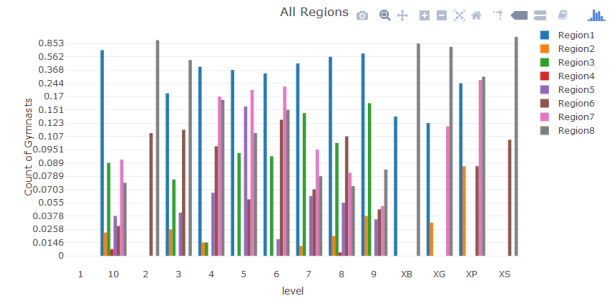


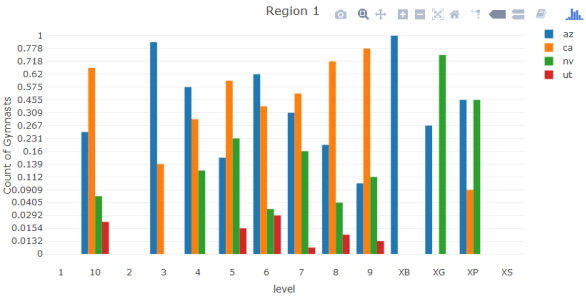


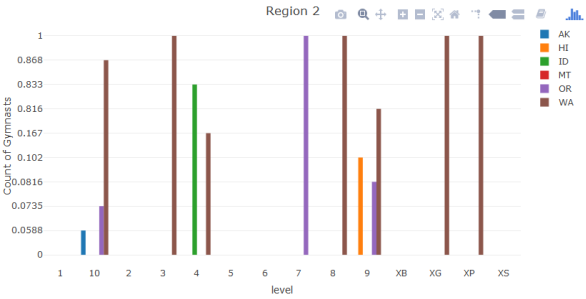


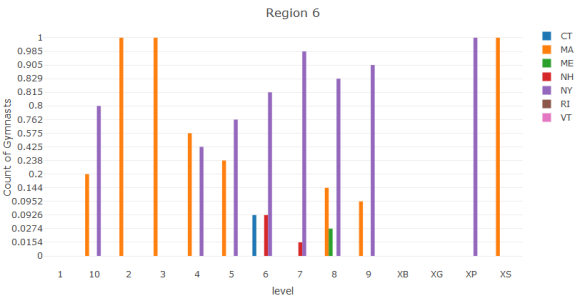


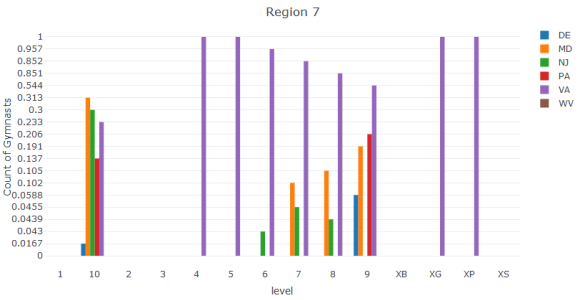
ppercentage of Gyms in each region and state with top100 gymnasts

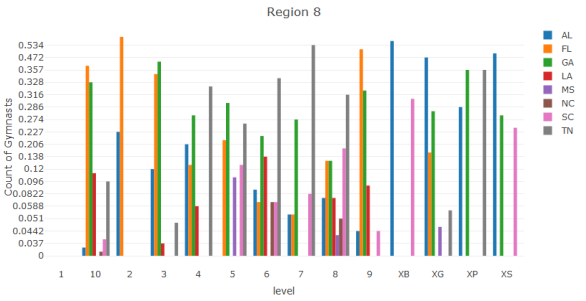






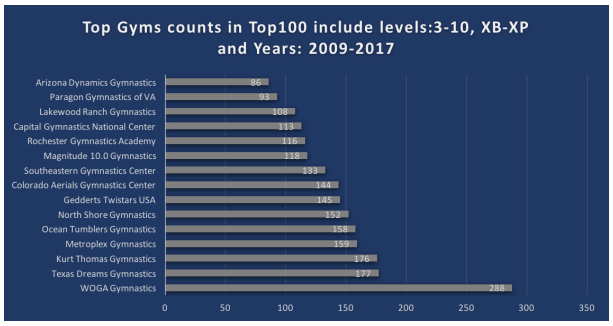




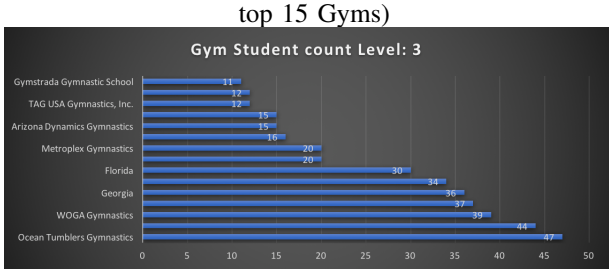


VI. RESULTS II

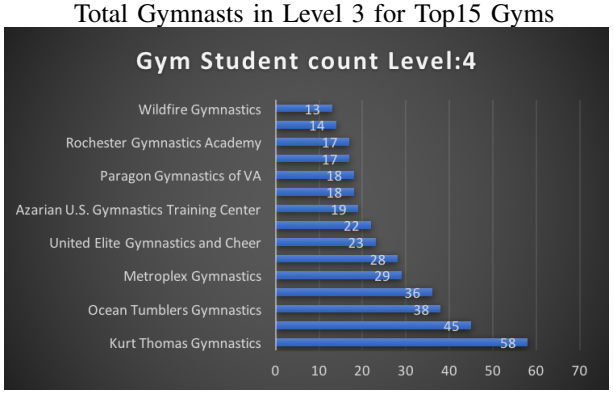
The questions we wanted to ask were for this set of results :Do particular Gyms impact a gymnast success?



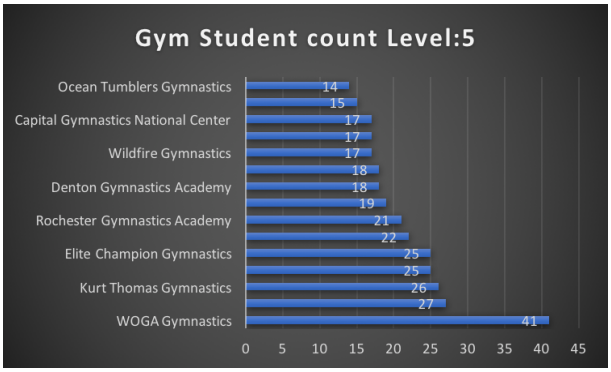
Total count of Gymnasts in Top100 for Gyms (this is the top 15 Gyms)



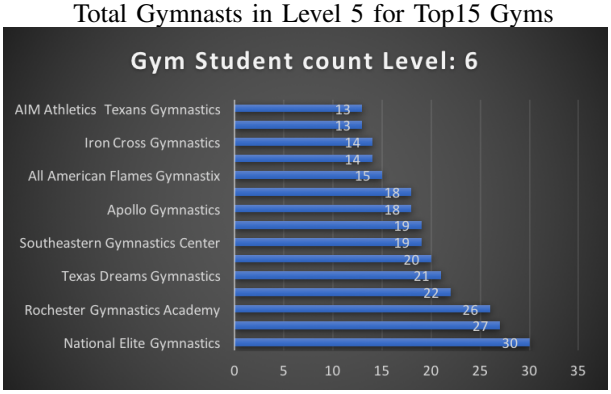
Total Gymnasts in Level 3 for Top15 Gyms



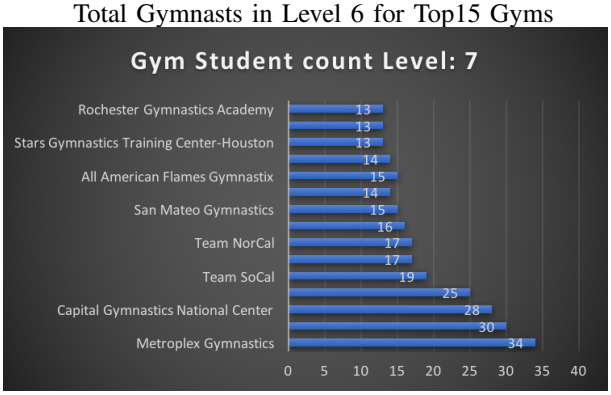
Total Gymnasts in Level 4 for Top15 Gyms



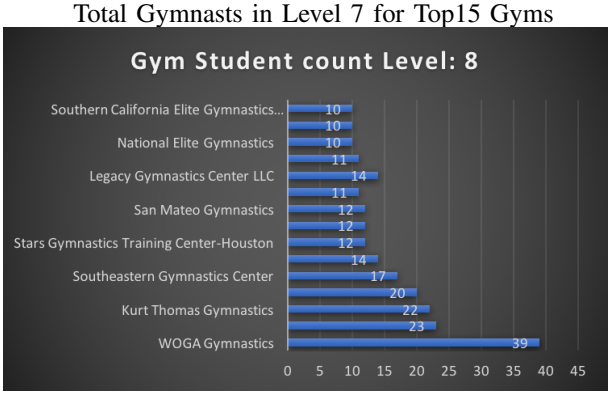
Total Gymnasts in Level 5 for Top15 Gyms



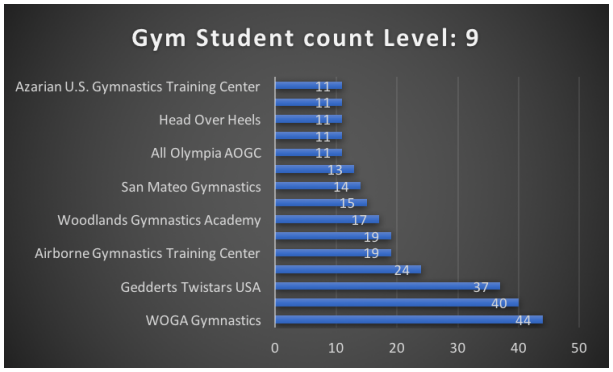
Total Gymnasts in Level 6 for Top15 Gyms



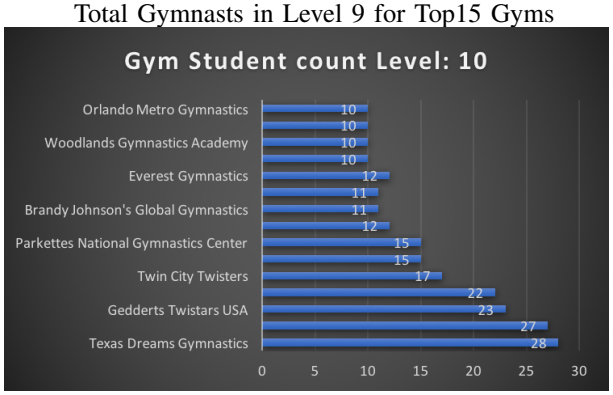
Total Gymnasts in Level 7 for Top15 Gyms



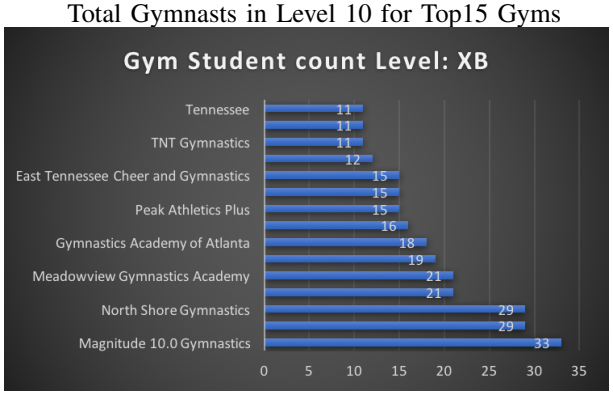
Total Gymnasts in Level 8 for Top15 Gyms



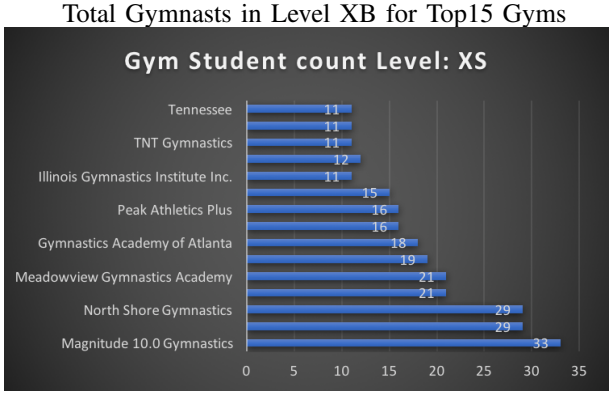
Total Gymnasts in Level 9 for Top15 Gyms



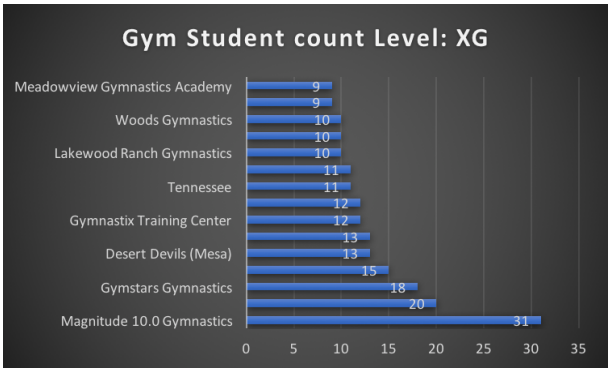
Total Gymnasts in Level 10 for Top15 Gyms



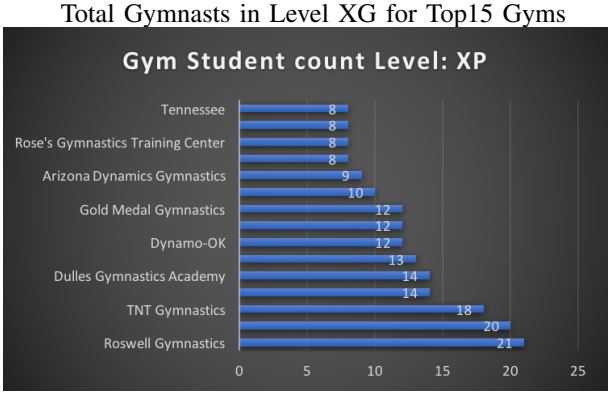
Total Gymnasts in Level XB for Top15 Gyms



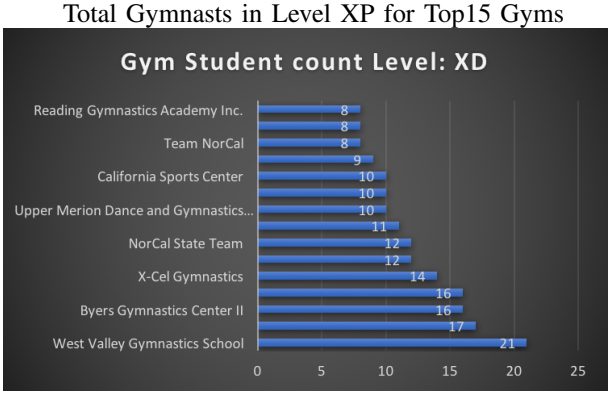
Total Gymnasts in Level XS for Top15 Gyms



Total Gymnasts in Level XG for Top15 Gyms



Total Gymnasts in Level XP for Top15 Gyms



Total Gymnasts in Level XD for Top15 Gyms

Connecting Databases

We were also able to connect the Meetme Score Online and My Meet Score. There was about 1583 matches between the two databases. We were not able to match with usa gym.

VII. Future Work

There is a difference between all round gymnast and event specialist. We want to be able to tell event with score is there is a way to differentiate the two or if all around score out ways the event specialist. We also want to see if it is possible to have more of a correlation based on overall rank of competitions over the year with improvement. We also want to see if we can make a program that is acceptable for all competitions sites. Though this seems to be not a plausible thing, there may be some sports with better defaults for site.

VIII. CONCLUSIONS

Result I: From what we can tell with top 100 the best Region to making it to level 10 and scoring well is Region 1. Region 8 Seems to be the best for overall getting top100s especially if your doing xcel. Early levels seem to be very subjective and this makes sense as children are really new to gymnastic in general.

Result II: It shows that WOGA Gymnastics have the highest number of Gymnasts that come in Top100 in Total result. Does that mean WOGA Gymnastics is one of the best Gym in USA? There is not enough data to prove it because the website did not provide enough information such as: total number of the Gymnasts for each level to see the success rate of the Gym.