Santiago Garcia

Fall 2015

I590-3417

H9 Final Project

Introduction

Crossfit is a fitness regime developed by Coach Greg Glassman. Over the years it has transformed into a competitive fitness sport that involves focusing on gymnastics, powerlifting, plyometrics, Olympic weightlifting among other exercises. While Crossfit is decentralized and not a specialized program, Crossfit focuses on increasing agility, stamina, , power, speed, coordination, balance, cardiovascular and respiratory endurance.

Crossfit has a set of competitions that are used in order to determine “the Fittest Man” or “the Fittest Woman” on Earth. To qualify for the Crossfit Games, athletes need to compete through various competitions in sequential manner that include the Crossfit Open, Regionals and finally the Crossfit Games.

Objective

My objective for this assignment is to take the data produced by the different athletes around the world to determine various relationships between those that compete in the Crossfit Regionals and those lucky athletes that win the Regionals competition that will compete in the Crossfit Games.

Tools for Analysis

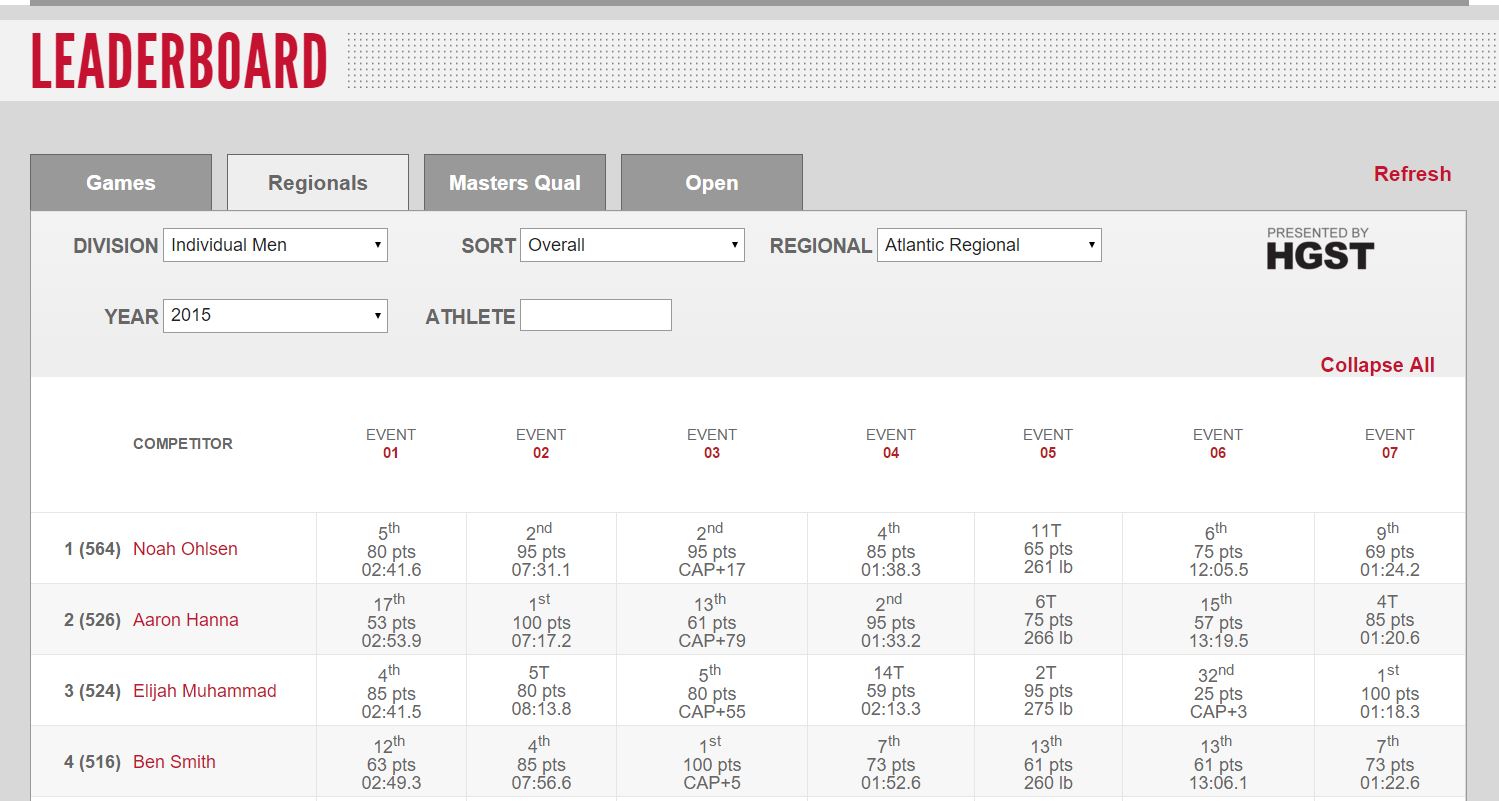
Canopy for Python Programming

Excel to store and maintain data in .csv files

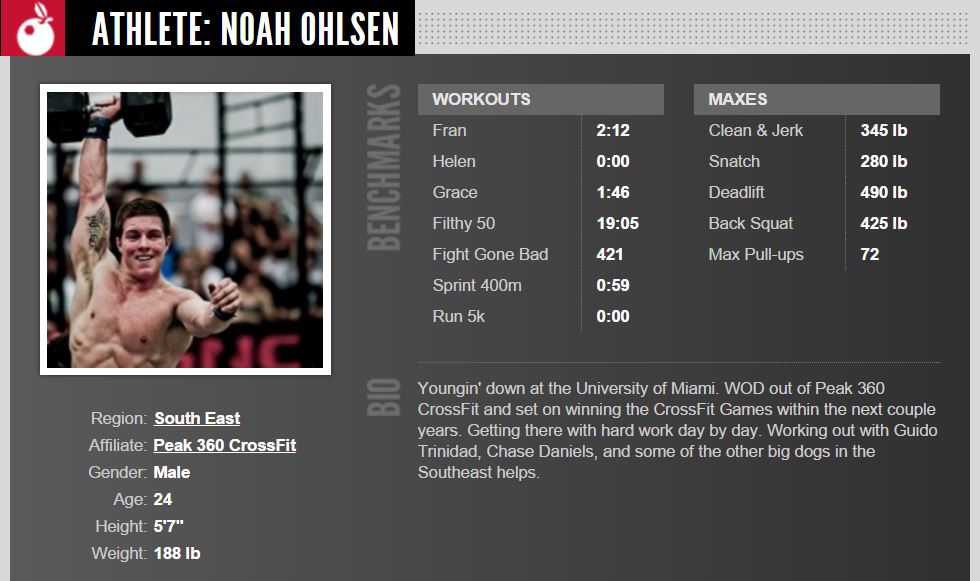
Pygal for visualization using Canopy and Python Programming

Obtaining Data for Analysis

Unfortunately, the company that owns Crossfit does not have a centralized repository where athlete information could easily be obtained. Any and all information can be obtained on the website via <http://games.crossfit.com/leaderboard>. The website contains a variety of tabs and drop downs. For this project, I focused on the Games and Regionals tabs. For the Regionals tabs, I made sure to only focus on the different regions. The Regional drop down menu contains 8 regions that correspond to different parts of the world.

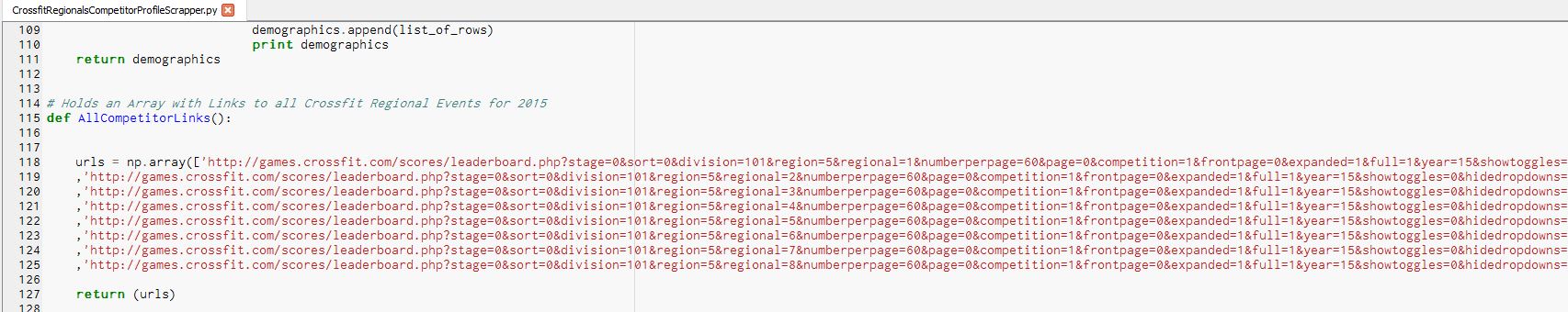


As mentioned above, in order to obtain the information I need, I needed to navigate through different drop downs. Each Regional area has information on the competitor, event, and times that a competitor took to finish an event. It is important to note that the information that is needed is that of only the competitor name. Each competitor name has a link that will allow the user to view that particular competitor’s profile information such as demographics, workout times, and maxes on certain weight lifting exercises.



The Web Scrapper

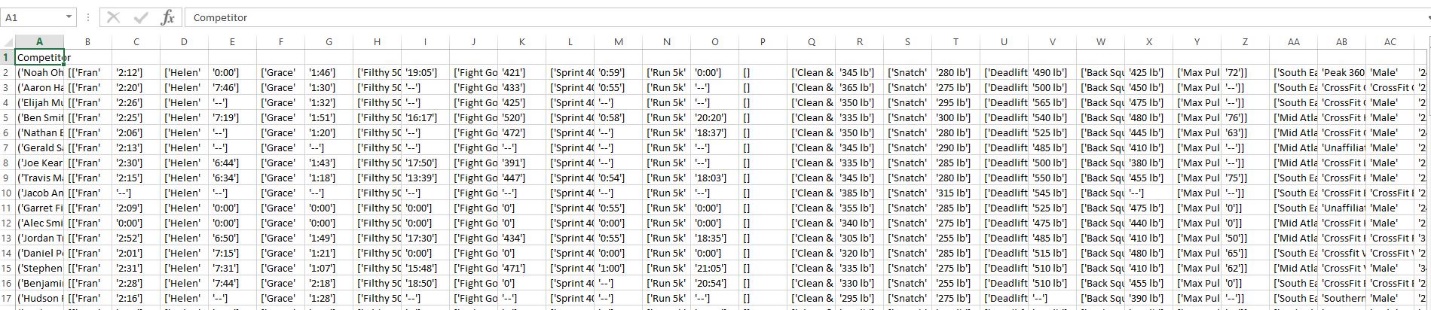
As noted before, the website that contains the information on competitors does not have a central repository. To the untrained eye, it would be easy just to copy and paste the information that is needed in order to analyze data. The problem with this approach is that process is very cumbersome, tedious and time consuming due to the fact that there are over 300 athletes that competed in the Crossfit Regionals. This would produce over 1,000 records that would need to be manually recorded. In order to speed up the process, I created a web scrapper in python that would automate the extraction of the data(CrossfitRegionalsCompetitorProfileScrapper.py). This web scrapper took the urls of each region (that I manually searched for) and searched for each competitor url (this happened programmatically) and put them in an array.



Once the profile urls were captured in an array, the program went through each and every profile extracting the data that was hardcoded in the html of the site. In addition, the same program created a .csv file (crossfitRegionals\_2015.csv) in which all the data was stored.

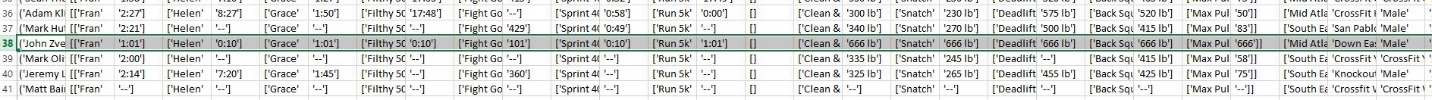
Cleaning the Data

Unfortunately, the data was not clean enough for me to process and analyze. The process that I created entered erroneous and sometimes invisible characters. I used the functions within Microsoft Excel and Microsoft Word to remove parenthesis, brackets, invisible tabs and any other erroneous characters by doing a search and replace. In addition, I had to add headers to each column that would be used later on in a different set of programs that would identify each column.



Once the data was cleaned, I saved the .csv file under a new name and called it “crossfitRegionals\_2015(Clean).csv”. In order to compare the Crossfit Regionals data to the Crossfit Games data, I repeated the entire process of developing a web scrapper and cleaning the data to create two more files “crossfitGamesProfileInfo\_2015.csv” and “crossfitGamesProfileInfo\_2015(Clean).csv”.

It is important to note that cleaning the data, while effective doesn’t always work. Sometimes with cleaning data, there could be inconsistencies, redundancy and/or missing data. Unfortunately for the parts that have missing data, I could not fill those cells with other data. To confirm, a handful of records, I proceeded to manually look for some of the data.



The above example illustrates an athlete with erroneous data such as having a “Clean & Jerk” at “666 lbs” or having the same weight for the “Snatch”, “Deadlift”, “Back Squat”, and “Max Pullups” respectively. While the above example is a bit strange, it is important to note that the data is not that quite erroneous due to the fact that the user purposely entered this data. Regardless, this is a great example of how data can be misrepresented, processed and later skew results.

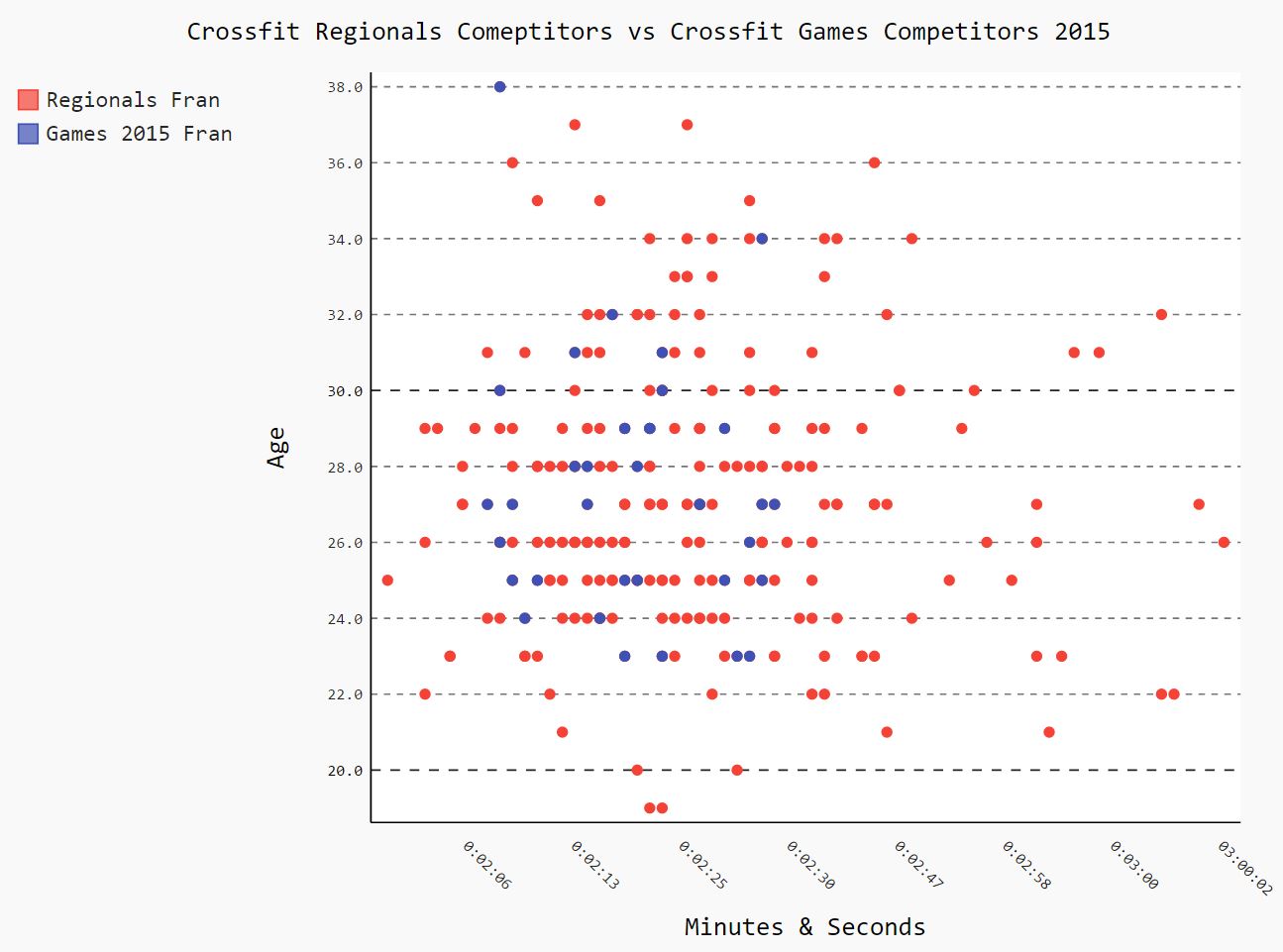
Processing the Data

Part of the Data Science process is to answer questions that have not yet been asked. But how do we do that? Well the easiest way is to start the data processing. After a brief check, the first place to start was with the columns that were mostly complete. The first column to process was the “Fran” column.

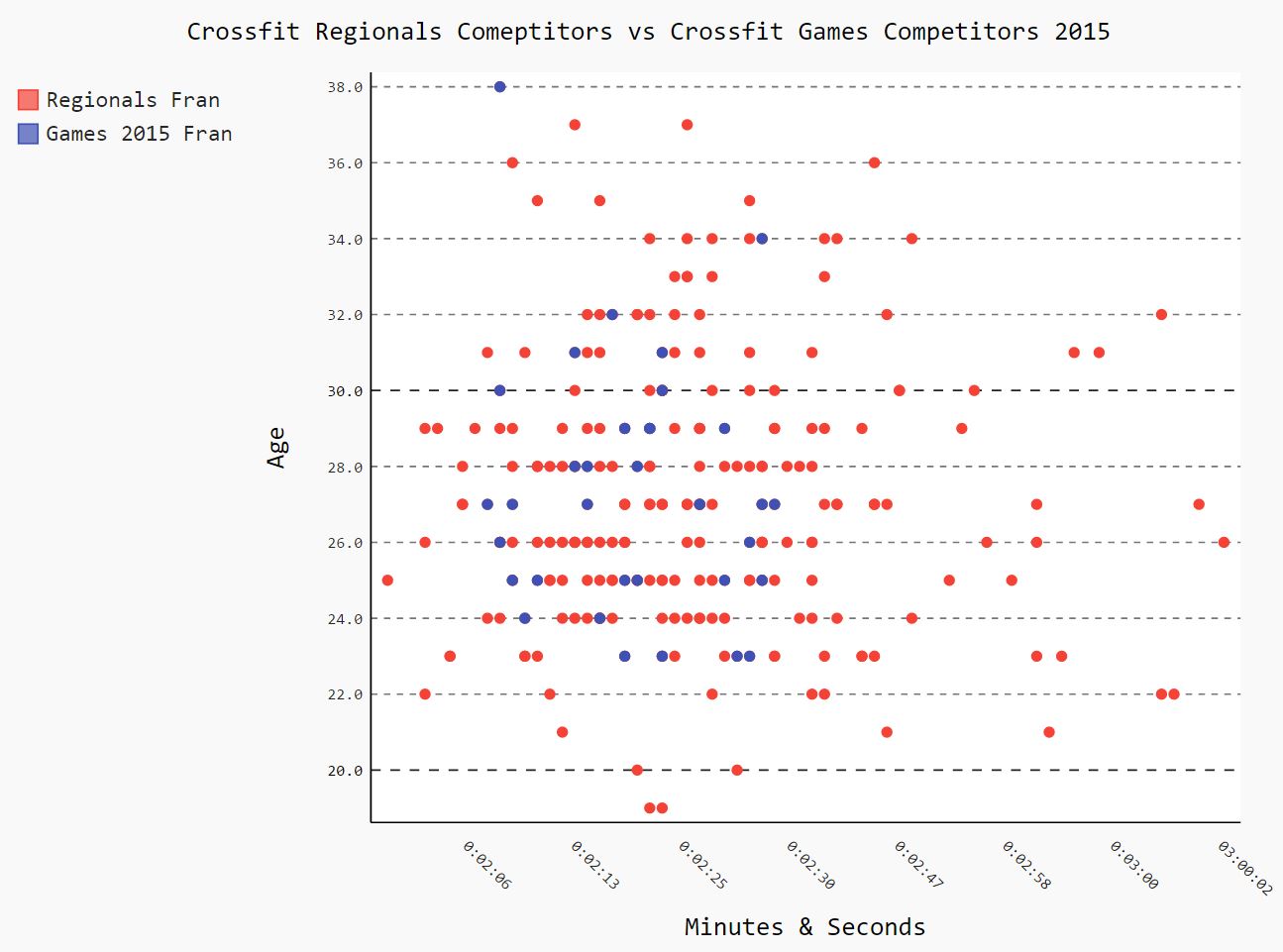
What is Fran?

Fran is a benchmark WOD (workout of the day) that consists of 3 rounds of 21-15-9 repetitions of 95 pound barbell thrusters, pull-ups all done for time. This workout is consistently used to monitor over time the increase in power and resistance for an athlete. This link takes you to a video where the competitor is doing Fran: <https://www.youtube.com/watch?v=pz9pXeLsmQk>

Regionals Fran vs Games Fran

To begin the analysis process, I created a program called CrossfitTimePlotting(Fran)\_2015.py to process data from two sources, “crossfitRegionals\_2015(Clean).csv” and “crossfitGamesProfileInfo\_2015(Clean).csv”. The data taken from these two files came from two columns, the first being the “Fran” column and the second being the “Age” column.

The above image illustrates Fran times in those that have competed in the Regionals vs the Games. The Regionals athletes are colored in red while the Games athletes are colored in blue. Based on the information above, in order to participate in the Crossfit Games, Regionals athletes should have a Fran time between 2 minutes 6 seconds and 2 minutes 27 seconds.



Using the same data, it is safe to assume that no athletes below the age of 23 participated in the Games while the oldest athlete was 38. Essentially there are very few athletes above the age of 30 that qualified for the Games while there are multiple athletes above the age of 30 that participated in Regionals.

Input files for creating the above scatter plot graph:

* crossfitRegoinals\_2015(Clean).csv
* crossfitGamesProfileInfo\_2015(Clean).csv
* CrossfitTimePlotting(Fran)\_2015.py

Output files for creating the above graph:

* CrossfitMaleFran\_2015.svg (Actual Scatter Plot Graph shown above)

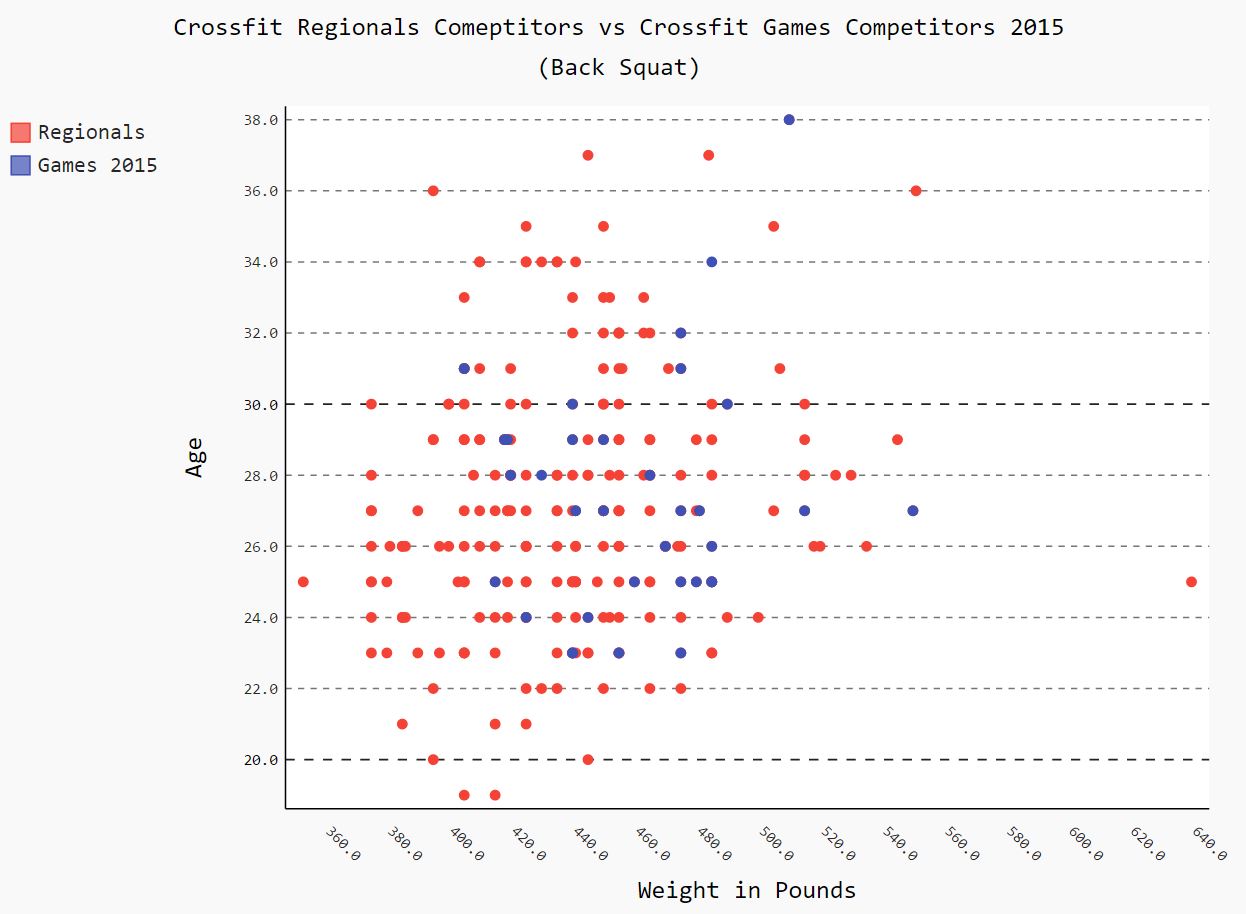
Intriguingly enough, the results from my first analysis created more questions than answers. The first question was why do the athletes under the age of 23 did not qualify for the Crossfit Games when some of those same athletes have better Fran times than those who participated in the Games. In order to answer the question, I built more scatter plot graphs based on the following criteria:

1. Max Back Squat/Age
2. Max Clean & Jerk/Age
3. Max Deadlift/Age
4. Max Snatch/Age
5. Competitor Weight/Age
6. Competitor Height/Age



The above image is a sample of some of the scatter plot graphs that were created with the previously mentioned criteria.

Max Back Squat



As the above image describes, those that qualified for the Crossfit Games had a Max Squat between the ranges of 400lbs. to 550lbs. while the bulk of the Games competitors lay between 400lbs and 490lbs. The competitors under the age of 23 fell who competed in Regionals fell within the range of the Crossfit Games competitors thus making the Back Squat not a factor of their failure to qualify for the Crossfit Games. The following link will take you to a video of Crossfit competitors working up to a Max Squat: <https://www.youtube.com/watch?v=_lmpUTcUTmM>

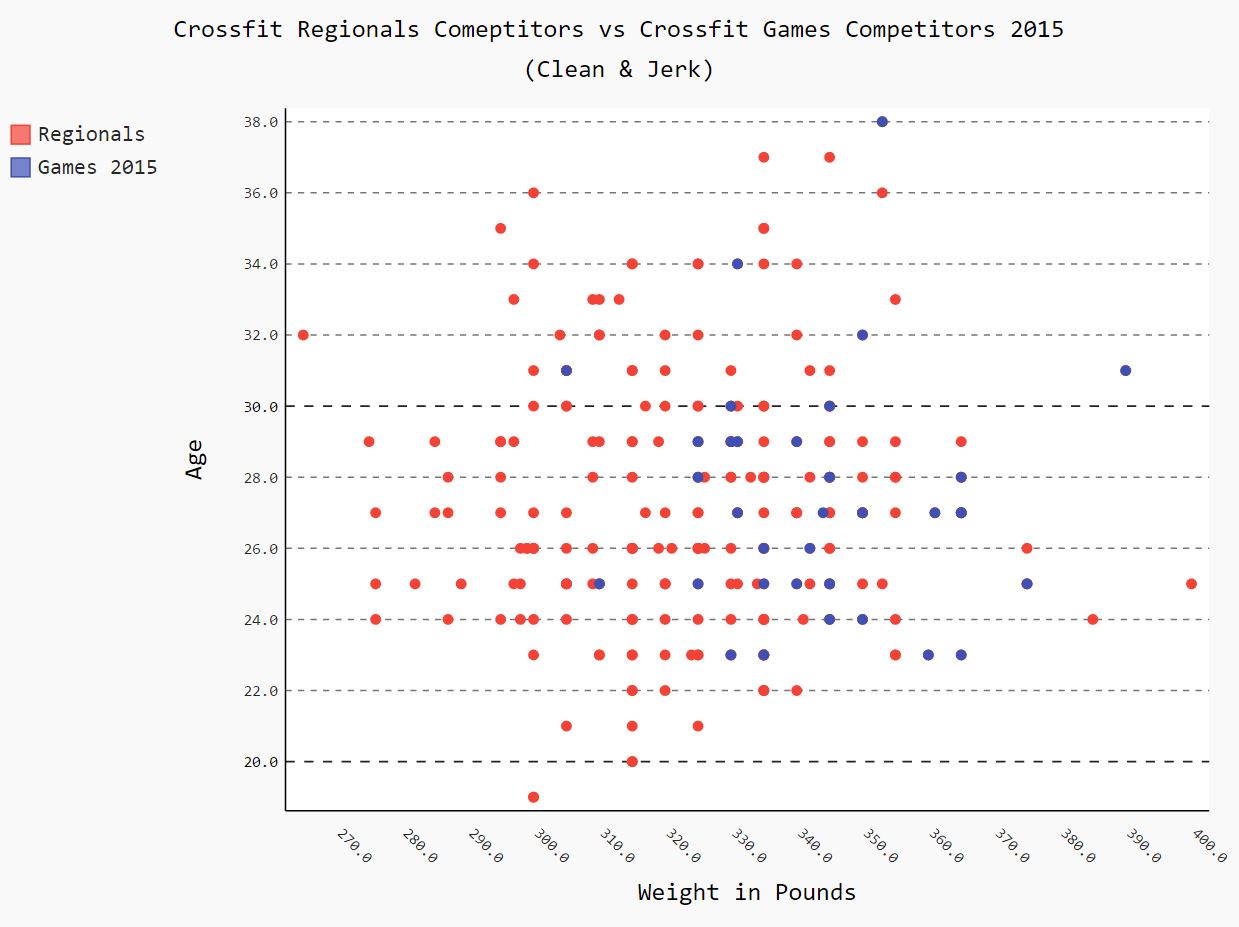
Input files for creating the above scatter plot graph:

* crossfitRegoinals\_2015(Clean).csv
* crossfitGamesProfileInfo\_2015(Clean).csv
* CrossfitWeightPlotting(BackSquat)\_2015.py

Output files for creating the above graph:

* CrossfitMaleBackSquat\_2015.svg (Actual Scatter Plot Graph shown above)

Clean & Jerk



As the above image describes, those that qualified for the Crossfit Games had a Max Clean & Jerk between the ranges of 305lbs. to 390lbs. while the bulk of the Games competitors lay between 320lbs. and 360lbs. The competitors under the age of 23 fell who competed in Regionals fell within the range of the Crossfit Games competitors thus making the Clean & Jerk just like the Back Squat not a factor of their failure to qualify for the Crossfit Games. The following link will take you to a video of the 2nd place winner of the Crossfit Games 2015 performing a max Clean & Jerk: <https://www.youtube.com/watch?v=zEztwTB5NhU>

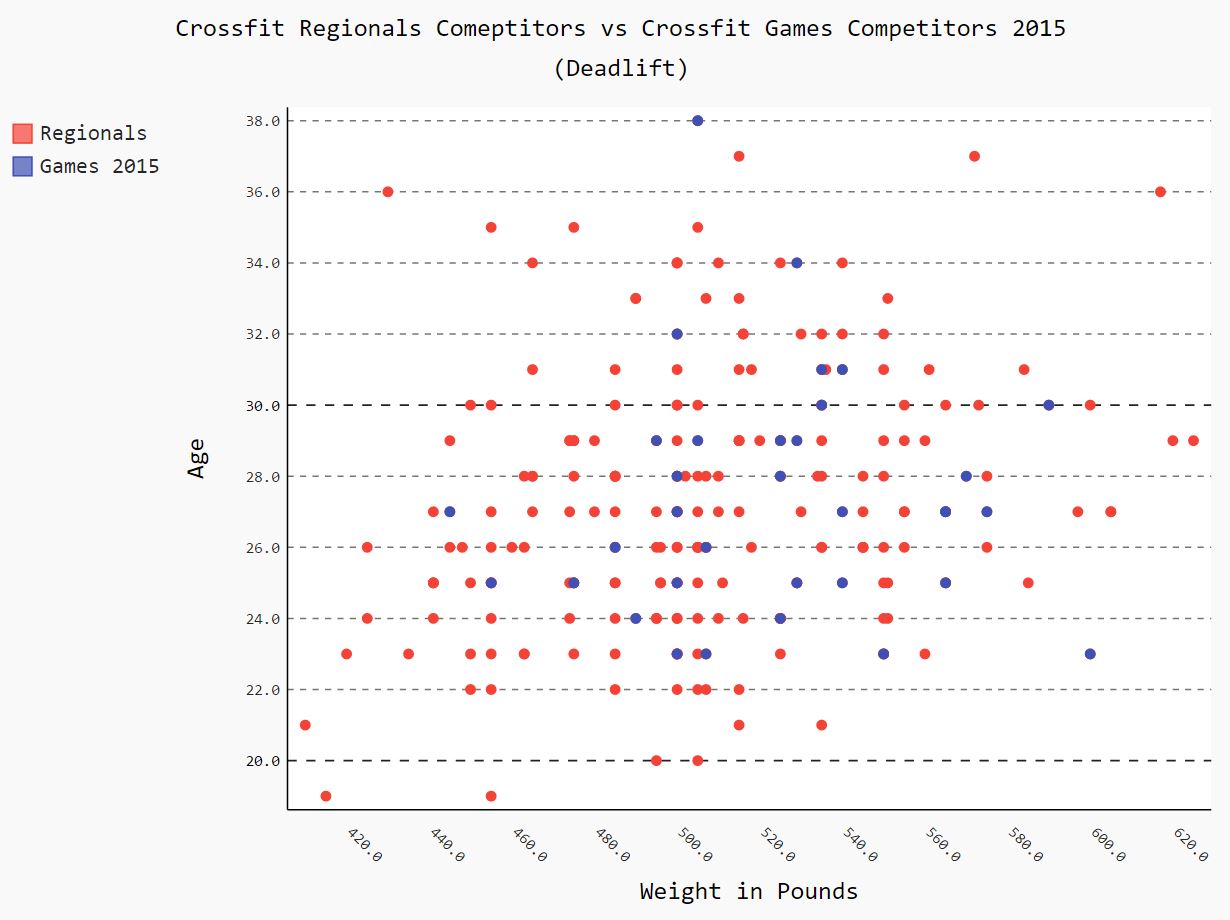
Input files for creating the above scatter plot graph:

* crossfitRegoinals\_2015(Clean).csv
* crossfitGamesProfileInfo\_2015(Clean).csv
* CrossfitWeightPlotting(CleanJerk)\_2015.py

Output files for creating the above graph:

* CrossfitMaleCleanJerk\_2015.svg (Actual Scatter Plot Graph shown above)

Max Deadlift

****

As the above image describes, those that qualified for the Crossfit Games had a Max Deadlift between the ranges of 440lbs. to 600lbs. Unlike the Squat and the Clean & Jerk, the ranges of the the plotted weights were a bit more scattered. And just like the previous scatter plot graphs, the competitors under the age of 23 fell who competed in Regionals fell within the range of the Crossfit Games competitors thus making this exercise as well not a factor of their failure to qualify for the Crossfit Games. The following link will take you to a video of a Crossfit athlete performing a Max Deadlift: <https://www.youtube.com/watch?v=boxNZPp1QOk>

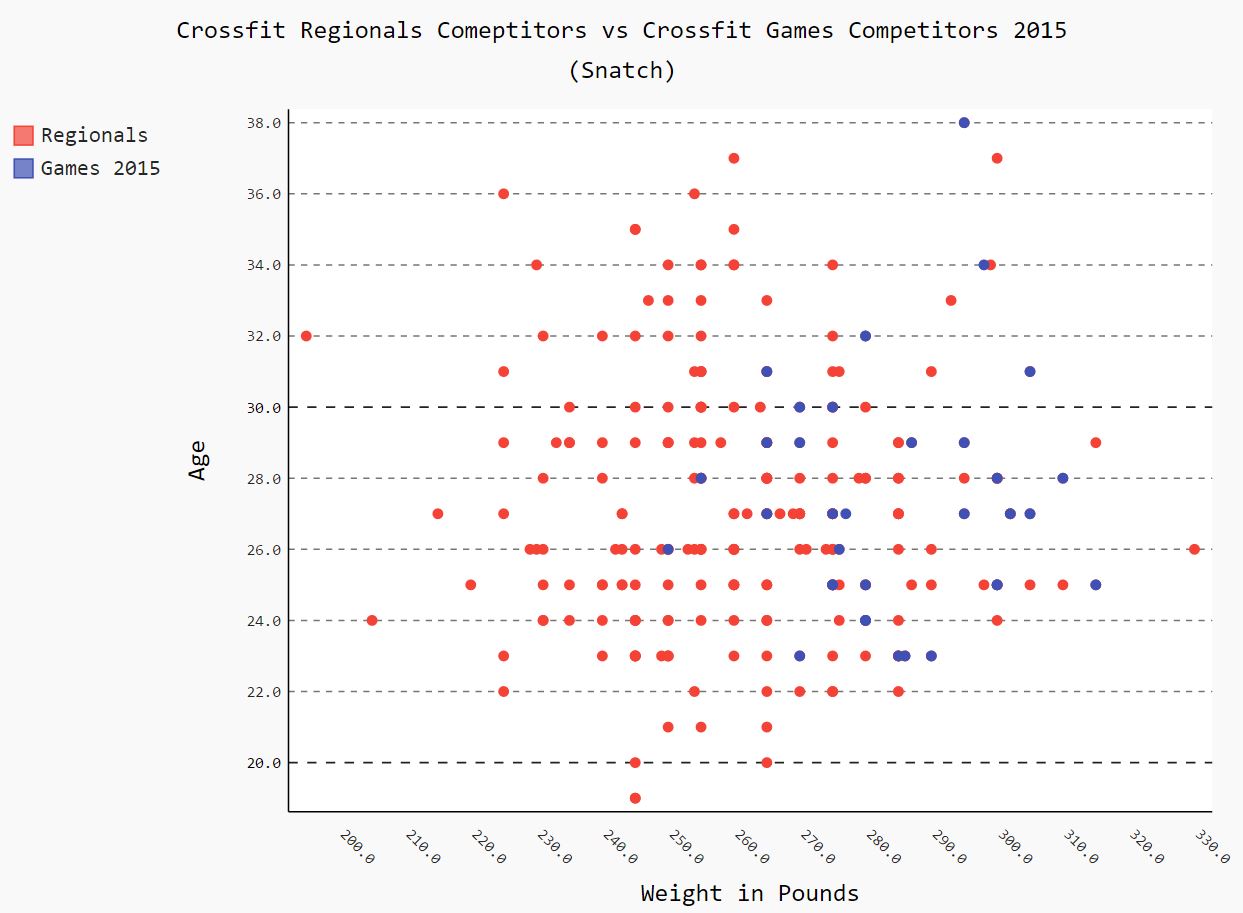
Input files for creating the above scatter plot graph:

* crossfitRegoinals\_2015(Clean).csv
* crossfitGamesProfileInfo\_2015(Clean).csv
* CrossfitWeightPlotting(Deadlift)\_2015.py

Output files for creating the above graph:

* CrossfitMaleDeadlift\_2015.svg (Actual Scatter Plot Graph shown above)

Max Snatch



As the above image describes, Crossfit Games competitors had a Max Snatch between the ranges of 250lbs. to 315lbs. Unlike the previous exercises before, there is finally a significant difference between Games and Regionals competitors. Games competitors lifted significantly more weight than those that didn’t qualify for the Games. Additionally, those competitors under the age of 23 fell significantly behind in their lifts. The bulk of these athletes under 23 fell between 220lbs. and 270lbs thus making the Max Snatch a possible factor in qualifying for the Games. The following link will take you to a video of the 2012, 2013, and 2014 Crossfit Games athlete Rich Froning performing a Max Deadlift: <https://www.youtube.com/watch?v=oWNSNtC6Kmc>

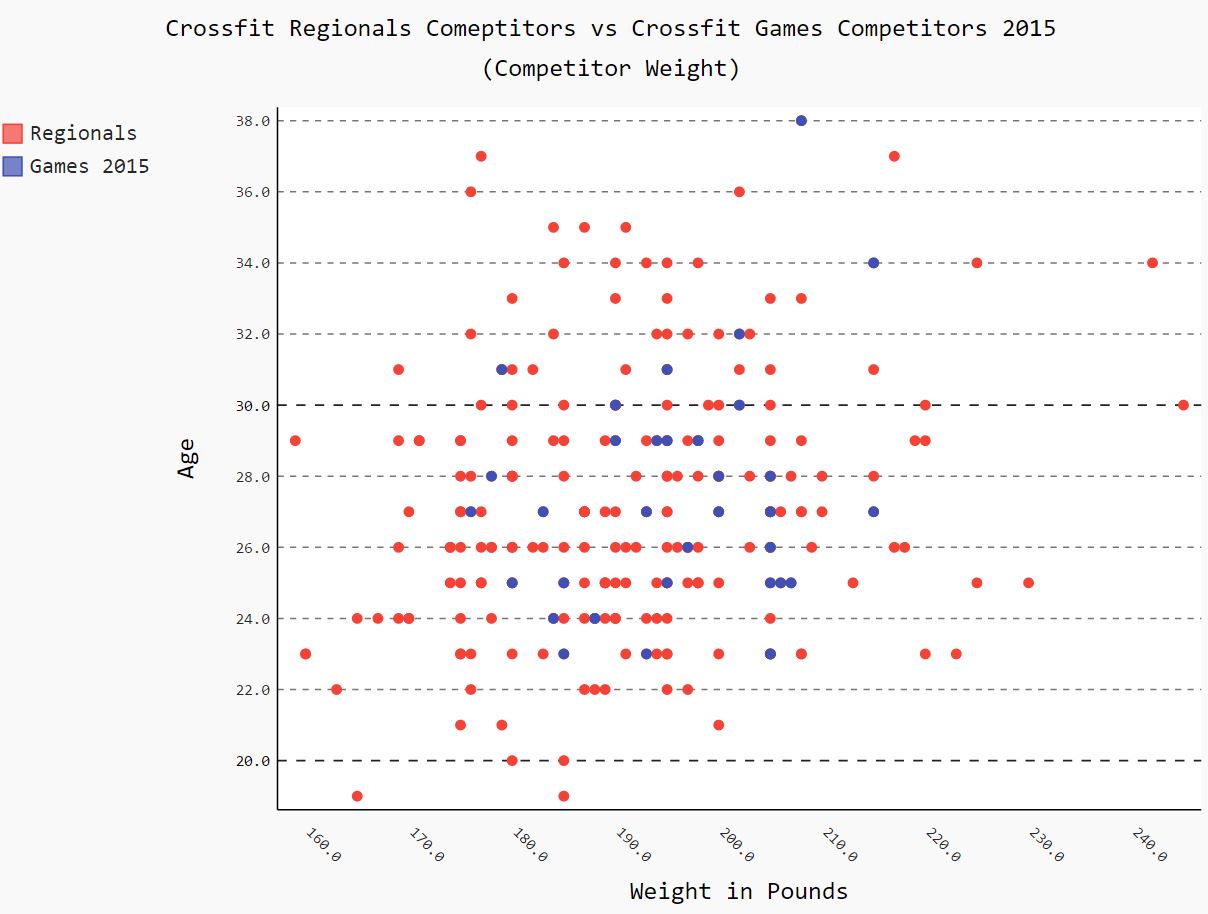
Input files for creating the above scatter plot graph:

* crossfitRegoinals\_2015(Clean).csv
* crossfitGamesProfileInfo\_2015(Clean).csv
* CrossfitWeightPlotting(Snatch)\_2015.py

Output files for creating the above graph:

* CrossfitMaleSnatch\_2015.svg (Actual Scatter Plot Graph shown above)

Competitor Age to Weight



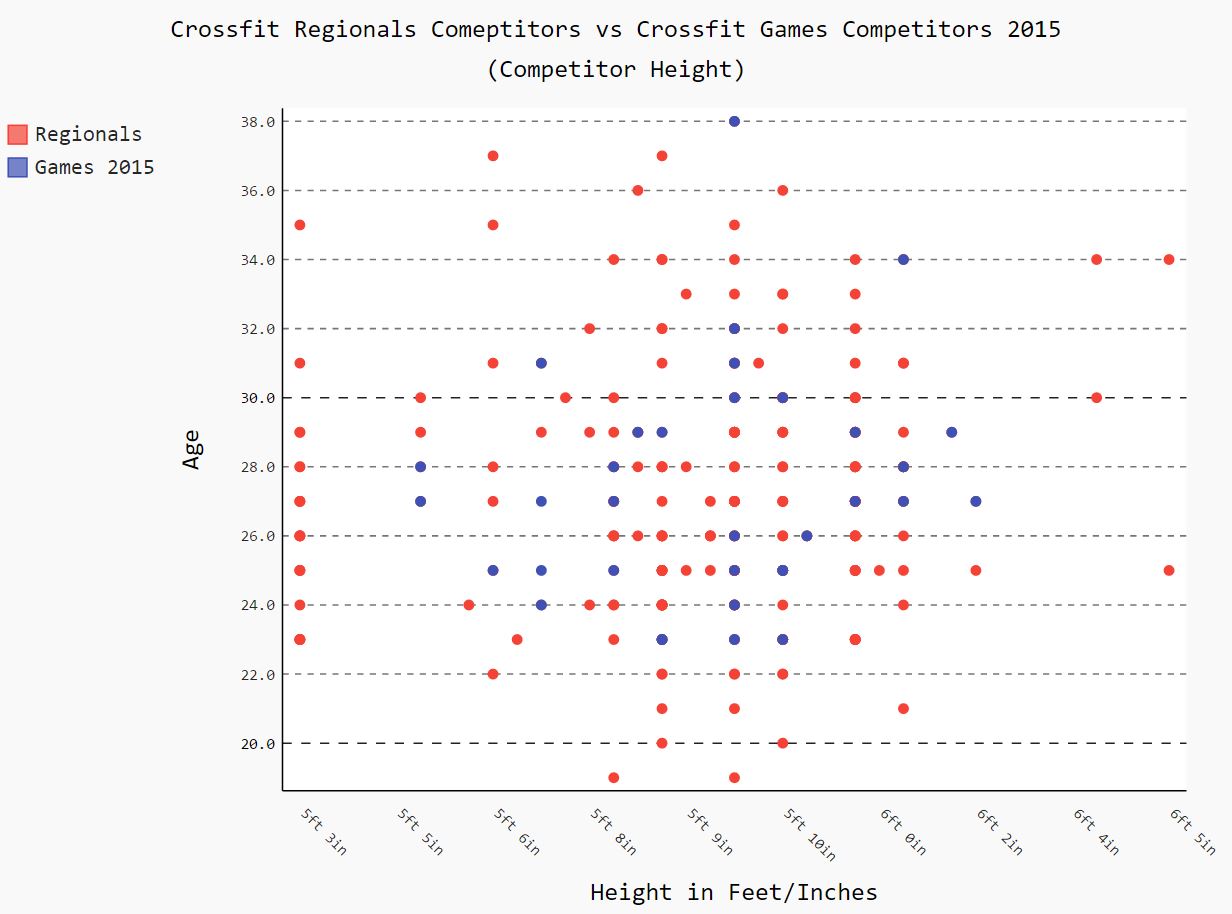
As the above image describes, Crossfit Games competitors weighed between the ranges of 175lbs. to 215lbs. The bulk of both Games and Regional athletes fell within these ranges including those athletes that were under the age of 23. Based on the image above, competitor weight was not a factor on whether a Regional competitor under the age of 23 qualified to compete in the Games.

Input files for creating the above scatter plot graph:

* crossfitRegoinals\_2015(Clean).csv
* crossfitGamesProfileInfo\_2015(Clean).csv
* CrossfitWeightPlotting(Weight)\_2015.py

Output files for creating the above graph:

* CrossfitMaleWeight\_2015.svg (Actual Scatter Plot Graph shown above)

Competitor Age to Height

As the above image describes, Crossfit Games competitors’ height ranges between 5ft. 5in and 6ft 2 in. Height did not play a significant factor in those athletes under the age of 23.

Input files for creating the above scatter plot graph:

* crossfitRegoinals\_2015(Clean).csv
* crossfitGamesProfileInfo\_2015(Clean).csv
* CrossfitHeightPlotting(Height)\_2015.py

Output files for creating the above graph:

* CrossfitMaleHeight\_2015.svg (Actual Scatter Plot Graph shown above)

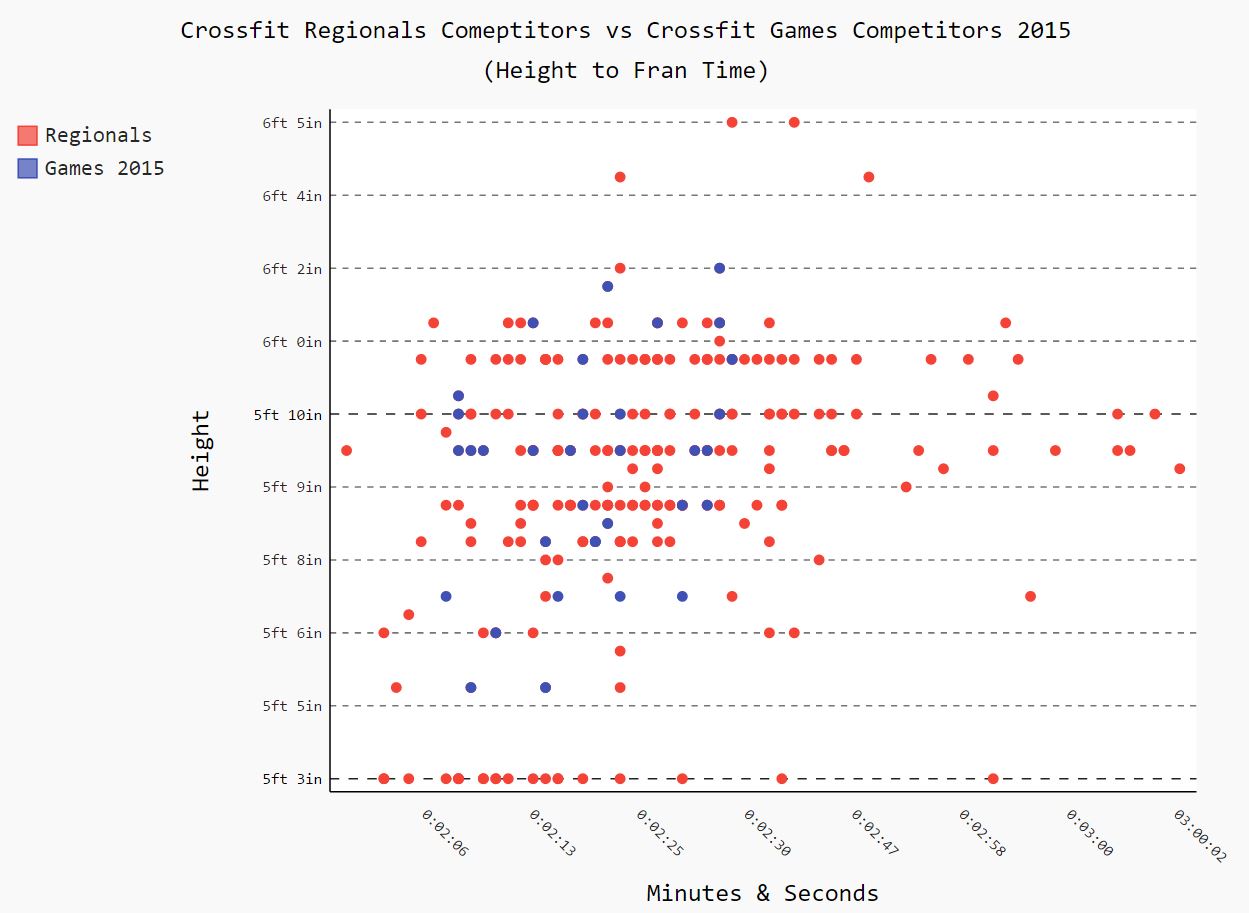
Conclusion

After analyzing multiple factors such as Fran Times, Max Back Squat, Max Clean & Jerk, Max Deadlift, Max Snatch, Competitor Weight and Competitor Height, we can determine on the data points that have been obtained that the possible factors that affect competitors under the age of 23 from competing in the Crossfit Games is their ability to Snatch above 250lbs. But is Age to Max Snatch the only factor that determines who qualifies the games. After analyzing the previous two graphs, it seems that there may be other factors. Could competitor height and weight also be another factor?

In the following section, we will be analyzing the exercise data against competitor height and weight ratio against the following:

* Fran Time
* Max Back Squat
* Max Clean & Jerk
* Max Deadlift
* Max Snatch

Height to Fran Time



As the above image describes, Crossfit Games competitors’ height ranges between 5ft. 5in and 6ft 2 in. While competitors in the 5ft 3in range had significant faster Fran times, it seems that there is no relationship between those competitors in smaller stature and Fran times. So why smaller athletes unable to qualify to compete in the Crossfit Games? Is it a matter of strength rather than an endurance exercise like Fran?

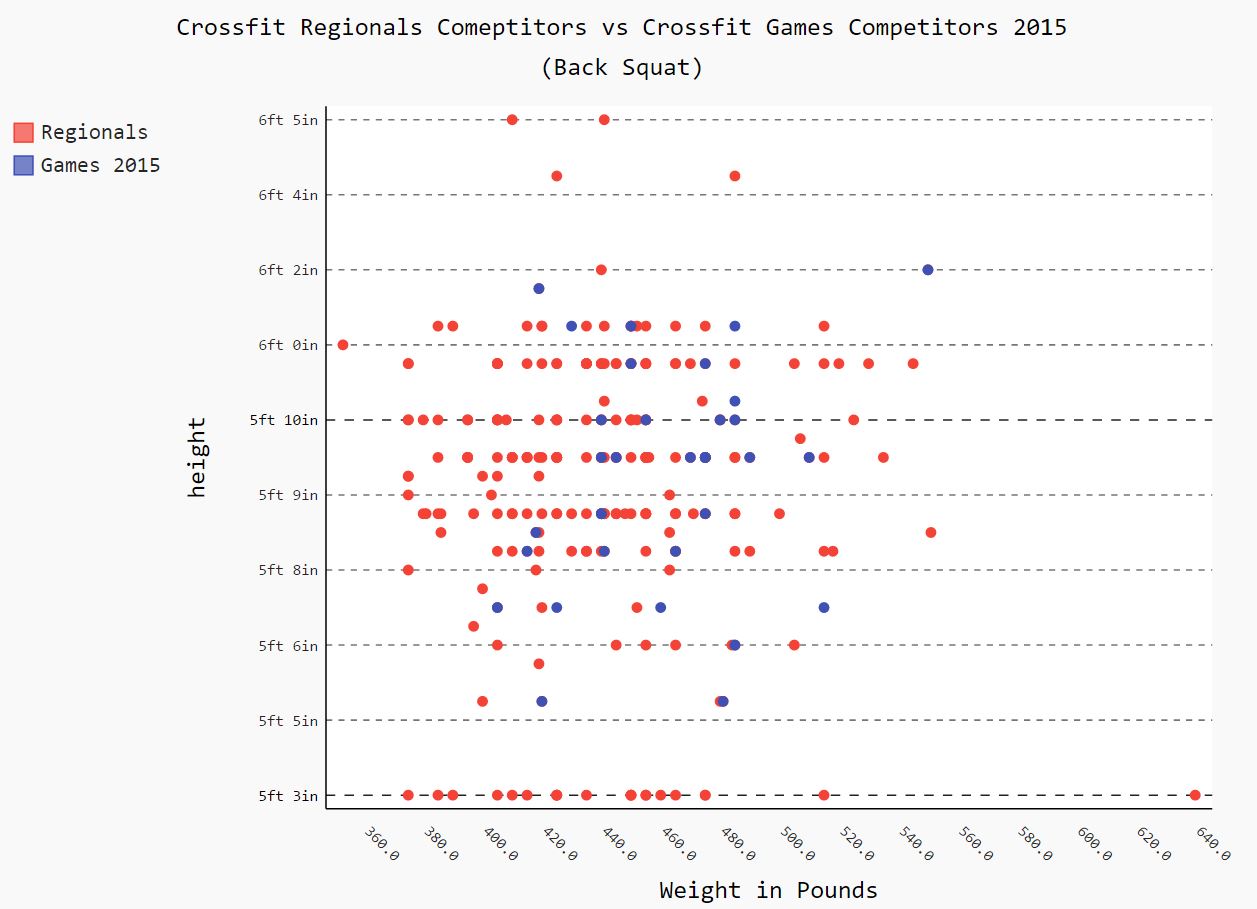
Input files for creating the above scatter plot graph:

* crossfitRegoinals\_2015(Clean).csv
* crossfitGamesProfileInfo\_2015(Clean).csv
* CrossfitHeightToFranPlotting\_2015.py

Output files for creating the above graph:

* CrossfitMaleHeightToFran\_2015.svg (Actual Scatter Plot Graph shown above)

Height to Max Squat



Once again height to Max BackSquat was not a factor as the majority of those competitors that fell in the 5ft 3in range were able to Max BackSquat as much or more than those that competed in the Games.

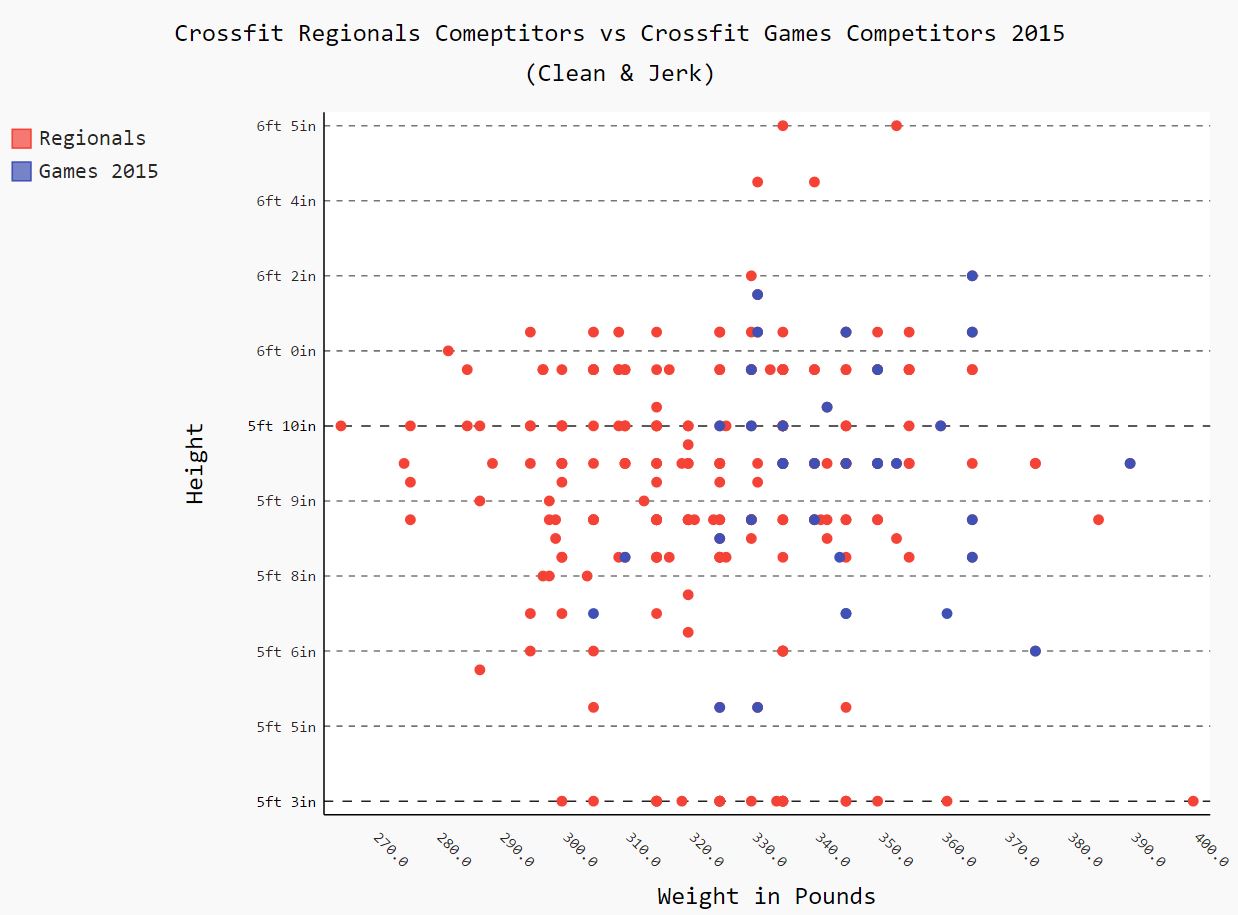
Input files for creating the above scatter plot graph:

* crossfitRegoinals\_2015(Clean).csv
* crossfitGamesProfileInfo\_2015(Clean).csv
* CrossfitHeightToSquatPlotting\_2015.py

Output files for creating the above graph:

* CrossfitMaleHeightToSquat\_2015.svg (Actual Scatter Plot Graph shown above)

Height to Max Clean & Jerk



Just as the Max BackSquat was not a factor in what caused competitors in the 5ft 3in range, the height to Clean & Jerk ratio was not a factor that determine this group from failing to compete in the Games. Consequently, an athlete at 5ft 3in was the competitor with the highest Clean & Jerk.

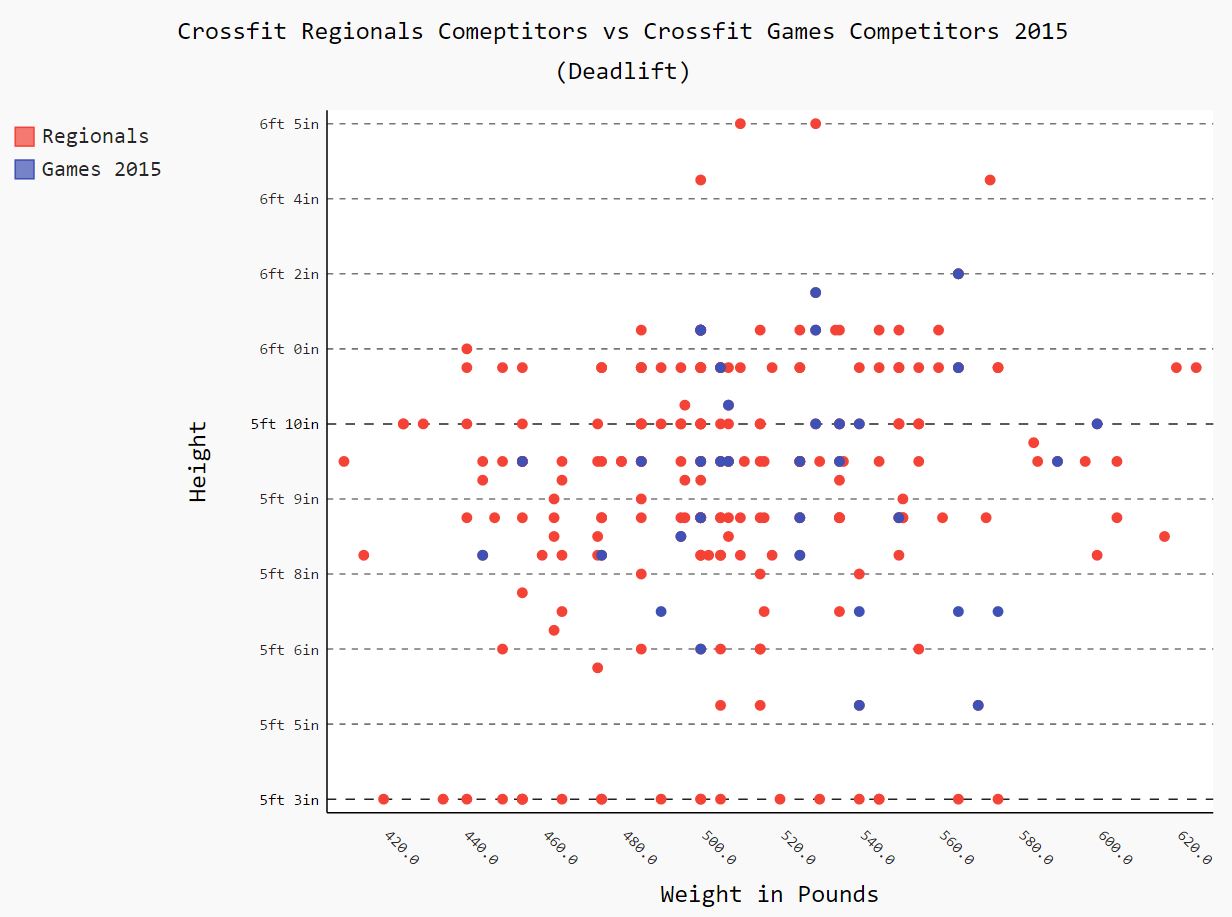
Input files for creating the above scatter plot graph:

* crossfitRegoinals\_2015(Clean).csv
* crossfitGamesProfileInfo\_2015(Clean).csv
* CrossfitHeightToCleanAndJerkPlotting\_2015.py

Output files for creating the above graph:

* CrossfitMaleHeightToCleanandJerk\_2015.svg (Actual Scatter Plot Graph shown above)

Height to Max Deadlift



Height had no significant effect on the amount of weight that is lifted during the deadlift. The majority of both Regionals and Game competitors fell within the range of 440lbs. to about 600lbs. Thus there was no significant relationship between height to Max Deadlift ratio.

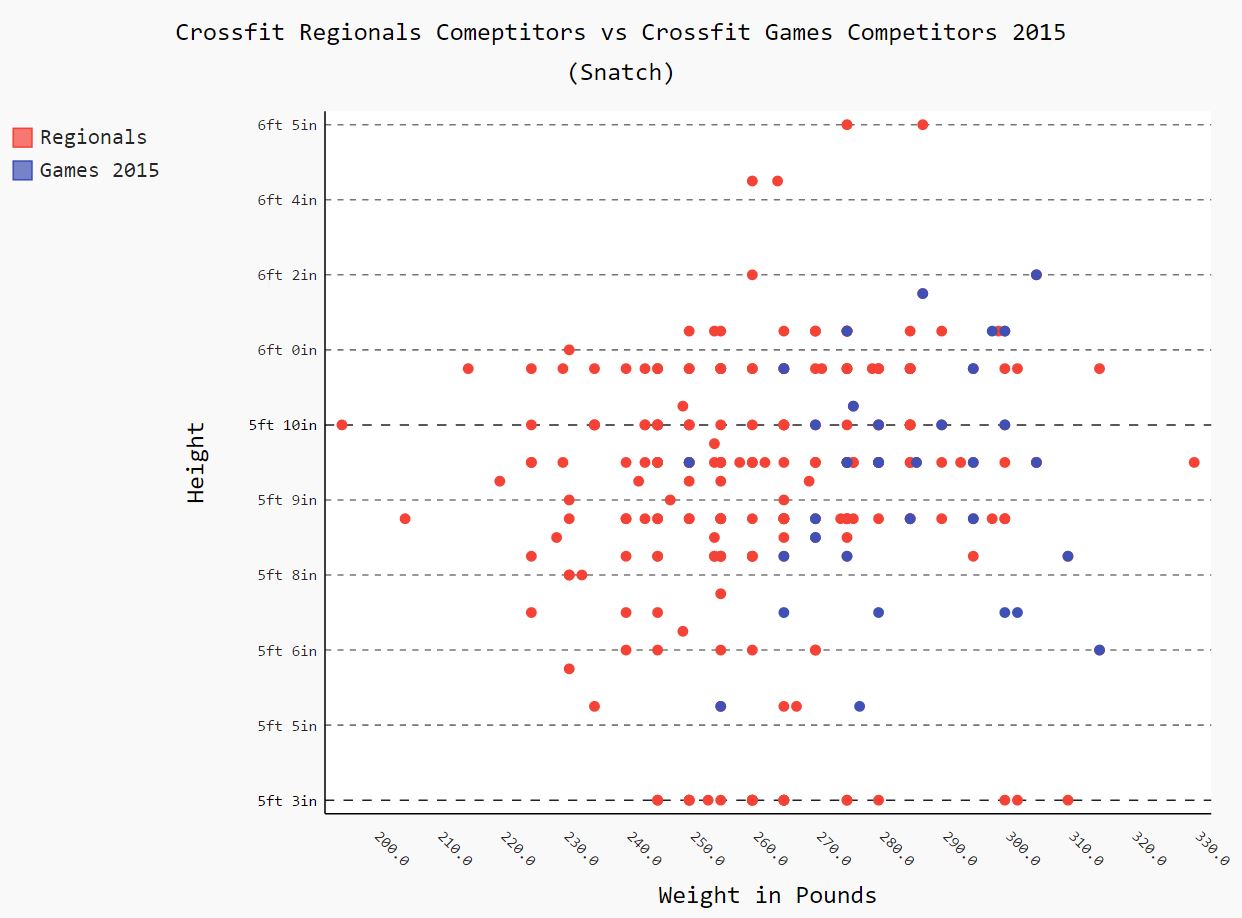
Input files for creating the above scatter plot graph:

* crossfitRegoinals\_2015(Clean).csv
* crossfitGamesProfileInfo\_2015(Clean).csv
* CrossfitHeightToDeadliftPlotting\_2015.py

Output files for creating the above graph:

* CrossfitMaleHeightToDeadlift\_2015.svg (Actual Scatter Plot Graph shown above)

Height to Max Snatch



While the Snatch range is between 250lbs. and 310lbs. for Games competitors, overall there was significantly more competitors below the 260lbs mark. Consequently there were a bit more competitors within the 5ft 3in range that were not able to lift more than 265lbs.

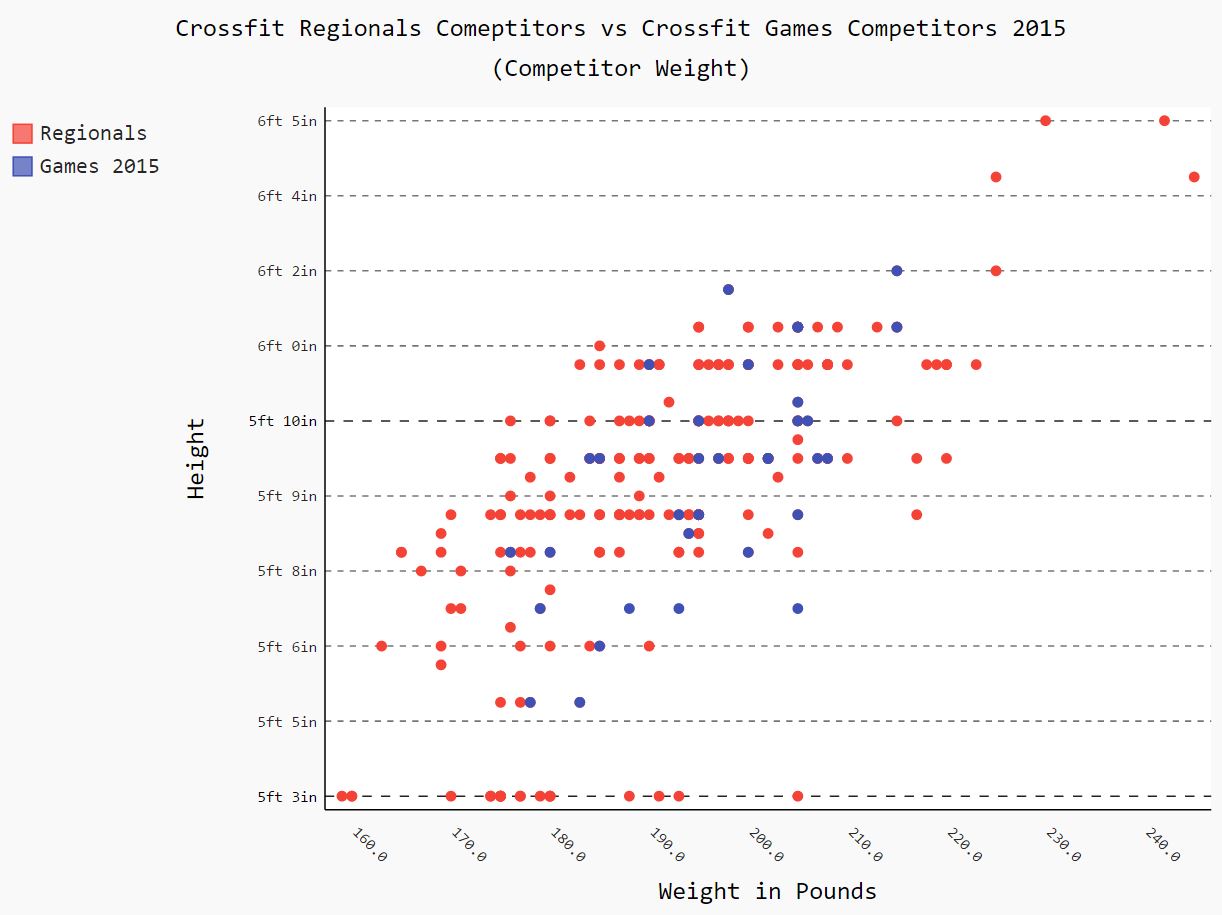
Input files for creating the above scatter plot graph:

* crossfitRegoinals\_2015(Clean).csv
* crossfitGamesProfileInfo\_2015(Clean).csv
* CrossfitHeightToSnatchPlotting\_2015.py

Output files for creating the above graph:

* CrossfitMaleHeightToSnatch\_2015.svg (Actual Scatter Plot Graph shown above)

Height to Weight



The weight range between Games competitors was between about 175lbs to 215lbs. Consequently there were a bit more competitors within the 5ft 3in range that were not able to lift weighted less than 175lbs.

Input files for creating the above scatter plot graph:

* crossfitRegoinals\_2015(Clean).csv
* crossfitGamesProfileInfo\_2015(Clean).csv
* CrossfitHeightToWeightPlotting\_2015.py

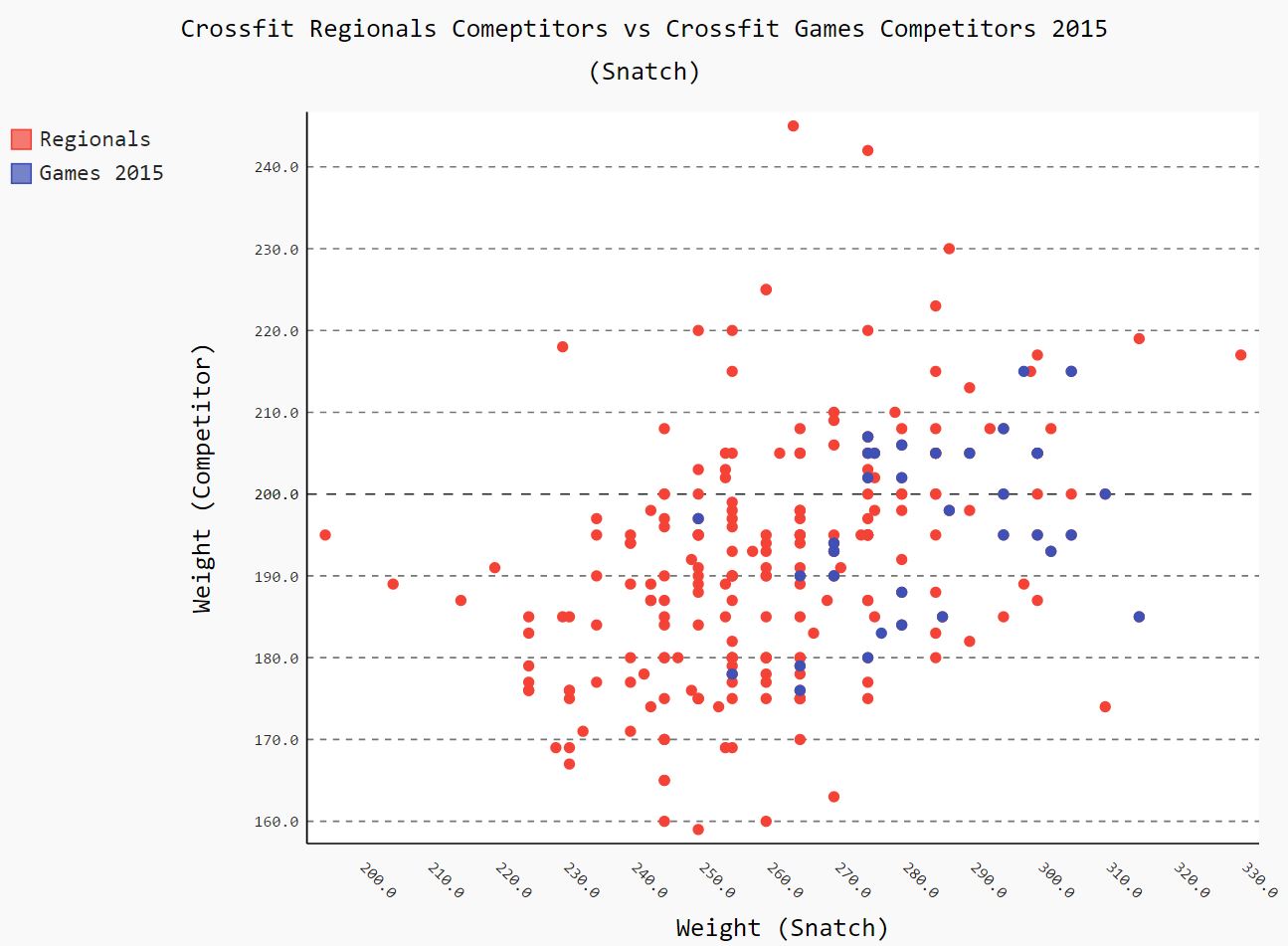
Output files for creating the above graph:

* CrossfitMaleHeightToWeight\_2015.svg (Actual Scatter Plot Graph shown above)

Conclusion

The results of this section seem a bit inconclusive as there seems to be no relationship between any of the above comparisons with the slight exception of the Max Snatch. Regardless, there is not enough evidence to support the reason why 5 ft 3in athletes did not compete in the Games.

Weight to Snatch



The image above illustrates that competitors in the Crossfit Games weighed a minimum of 180lbs while lifting a minimum Max Snatch of 250lbs. Based on the graph, it seems that there is a relationship between competitor weight and the ability to lift a Max Snatch that will allow competitors to enter into the CrossFit Games. The more they weight the greater their ability to lift.

Input files for creating the above scatter plot graph:

* crossfitRegoinals\_2015(Clean).csv
* crossfitGamesProfileInfo\_2015(Clean).csv
* CrossfitWeightToSnatchPlotting\_2015.py

Output files for creating the above graph:

* CrossfitMaleWeightToSnatch\_2015.svg (Actual Scatter Plot Graph shown above)

Overall Conclusion

After carefully review each and every graph, we can easily determine that in order to participate in the CrossFit games, competitors should be over the height of 5 ft. 3in but no taller than 6ft 2in. In addition to height being a factor, the Max Snatch is a significant factor in entering the Games. Competitors should be able to lift over 250lbs. Other significant factors is the competitor weight to Max Snatch ratio. In order for a competitor to lift over 250lbs. competitors should weight over 170lbs. While there are many other factors that may determine whether a Regionals competitor enters the CrossFit Games, there was not any other significant evidence or data that could be pulled in order to make further opinions.

Instructions

1. Install Canopy onto Windows/Mac device
2. Pip install Pygal for rendering graphs.
3. Go to Package Manager in Canopy and install the BeautifulSoup package, the numpy package, and the requests package
4. You will need to run the first program to extract the Regionals data from the various CrossFit Games links: CrossfitRegionalsCompetitorProfileScrapper.py. This should take approx. 1 minute 30 seconds to extract all the data. This should be dependent on your machine and internet connection
5. You will need to run the second program to extract the Crossfit Games 2015 data from the various CrossFit Games links: CrossfitGames2015CompetitorProfileScrapper.py
6. Once both of these files are created, you need to search for those files as they are produced locally and is dependent on whether you are on a Windows or Mac machine. A simple search should do. Files created are crossfitRegionals\_2015.csv and crossfitGamesProfileInfo\_2015.csv
7. Once you have located these files you can begin cleaning the files by removing erroneous characters and spaces. For the purpose of this project/class, this step can be skipped as two clean files are provided. These files are crossfitRegionals\_2015(Clean).csv and crossfitGamesProfileInfo\_2015(Clean).csv
8. These two files MUST be included in the same directory as the two previous files that were initially created. The creation of the visualized data will not work if the files are not in the correct directories.
9. Once the (Clean) files are in the proper directory, you can begin running the following programs:
   1. CrossfitTimePlotting(Fran).py
   2. CrossfitWeightPlotting(BackSquat)\_2015.py
   3. CrossfitWeightPlotting(CleanJerk)\_2015.py
   4. CrossfitWeightPlotting(Deadlift)\_2015.py
   5. CrossfitWeightPlotting(Snatch)\_2015.py
   6. CrossfitWeightPlotting(Weight)\_2015.py
   7. CrossfitHeightPlotting(Height)\_2015.py
   8. CrossfitHeightToFranPlotting\_2015.py
   9. CrossfitHeightToSquatPlotting\_2015.py
   10. CrossfitHeightToCleanAndJerkPlotting\_2015.py
   11. CrossfitHeightToDeadliftPlotting\_2015.py
   12. CrossfitHeightToSnatchPlotting\_2015.py
   13. CrossfitHeightToWeightPlotting\_2015.py
   14. CrossfitWeightToSnatchPlotting\_2015.py
10. Once you have ran the previous files, you can begin looking at the results using the following files:
    1. CrossfitMaleBackSquat\_2015.svg
    2. CrossfitMaleFran\_2015.svg
    3. CrossfitMaleBackSquat\_2015.svg
    4. CrossfitMaleCleanJerk\_2015.svg
    5. CrossfitMaleDeadlift\_2015.svg
    6. CrossfitMaleSnatch\_2015.svg
    7. CrossfitMaleWeight\_2015.svg
    8. CrossfitMaleHeight\_2015.svg
    9. CrossfitMaleHeightToFran\_2015.svg
    10. CrossfitMaleHeightToSquat\_2015.svg
    11. CrossfitMaleHeightToCleanandJerk\_2015.svg
    12. CrossfitMaleHeightToDeadlift\_2015.svg
    13. CrossfitMaleHeightToSnatch\_2015.svg
    14. CrossfitMaleHeightToWeight\_2015.svg
    15. CrossfitMaleWeightToSnatch\_2015.svg
11. Note: because these files are .svg files, if they are run on a Windows machine, they take significantly longer to render, aprox 1 minute. If these files are run on a Mac machine, the files render almost immediately
12. The source code can be found on <https://github.com/santgarc/Fall2015-1590-3417>