Data Analysis of the WKF Rankings Database

Ross Leeper

Executive Summary

The WKF Rankings Database contains all the WKF since the rankings system was implemented. In conjunction with data-analysis techniques, the database was used to answer questions on what countries are top tier along with the composition of that success, if performances at younger age categories give indications of future successes and how important frequent competition attendance is for performance over a competitors career. The analysis was aimed exclusively at Kumite events.

Based on a competition win-rate of greater than two wins per tournament and having won at least one gold medal at the last two world championships there a seven countries that could be considered top-tier. These are Japan, France, Iran, Egypt, Italy, Turkey and Germany. Japan is comfortably the best performing all-round country with France and Turkey being also being fairly consistent. Iran is much more successful in male categories as well as the medium and upwards weights.

Data-mining analysis using k-nearest neighbour techniques was unable to identify future medallists at senior level by using competition results for competitors at cadet, junior or under-21 events. This is due to data from these competitions not providing distinct features that can be found using K-NN data mining.

A medium to strong correlation was found when comparing frequency of attendance of all tournaments to medals won at either world championship or continental championship level. However there many outliers in this comparison so the recommendation that more competition attendance leads to more medals at high levels should not be considered a sole reason for lack of success in any setting.

All recommendations and findings from the database most not be considered as absolutes as there are many variables that will not have been possible to capture that have influence over the results of all WKF karate competitions.

Introduction

Karate is a martial art that it practised worldwide in 192 countries with an estimated 100 million practitioners (Aina, 2017). As a result of its popularity it has become worldwide and Olympic sport, with an accompanying governing body, the World Karate Federation (WKF, 2021a), that manages the sport. Part of the organisation of the sport involves maintaining records of tournament results for each competitor and a structured ranking system so only the most elite athletes compete at the top events. Using these tournament results collectively, this analysis will seek to answer four questions:

1. Are the perceived top tier countries actually top tier?
2. Are the top tier countries identified in Question 1 specialists in a specific weight or sex?
3. Does success in younger (Junior/Cadet/Under 21s) sections reliably lead to success in senior categories at the highest level?
4. Does frequency of attendance of small tournaments correlate with world or continental federation success?

These questions will be answered in turn with any findings or recommendations being found within the same section as the analysis pipeline.

This report will not require specialised knowledge from outwith the report.

Data Source

The dataset for this analysis was obtained with permission from SportData from the WKF Rankings website (Sportdata, 2021) by using a web scraper, an automated program for copying only required information from a website, and placing it within a new database format. This reformatting allowed for more complex analysis techniques to be performed on the dataset. The web scraper ran from 26th Match 2021 through to 1st April 2021.

The format of the information on each competitors entry into a competition was presented with ten data points: ranking country; ranking competitor ID; date of the event; name of the event; type of event; category entered; event factor of the competition; finishing rank of the competitor at the event; number of matches won by the competitor; WKF ranking points awarded to the competitor.

Analysis

It must be noted, on March 11th 2020 a global pandemic was declared by the World Health Organisation (WHO) in light of the spread of the COVID-19 coronavirus. This subsequently caused many countries to impose travel restrictions and cancel events to prevent the spread.

As a result of this, data from events after March 11th 2020 has been removed from all analysis. It cannot be guaranteed that it is consistent with the rest of the data from periods with no travel restrictions and athletes not having the same preparation, tournament access or training facilities.

WKF competitions have two main disciplines known as *Kata* (forms) and *Kumite* (sparring). Both hold events at the same tournaments but there is almost no overlap between competitors and these forms of competition. Although Kata results would form a fractional subset of the data due to the different weight categories in Kumite competition, technical differences between Kata and Kumite means that these results should be considered separately. Therefore the scope of this analysis will solely be on Kumite results.

Are the perceived top tier countries actually top tier?

From initial consultation with stakeholders in the karate domain, the countries that are perceived as top tier are Japan, France, Turkey, Azerbaijan, Iran and Egypt.

The results of the three most recent Senior WKF World Championships from 2018, 2016 and 2014 were identified as the best measure for top tier as they gain the highest event factor score and have results that are still relevant to competitors and coaches today.

The counties will be viewed as a whole rather than a selective of the most successful competitors, therefore all results of all competitors entered in an event as a member of a country will be included.

All the results of each competitor were grouped together by country and a mean average of the matches won per event for each country was calculated. The countries that had an average matches won of greater than or equal to two and had a won at least one 1st place finish are shown in Figure 1.

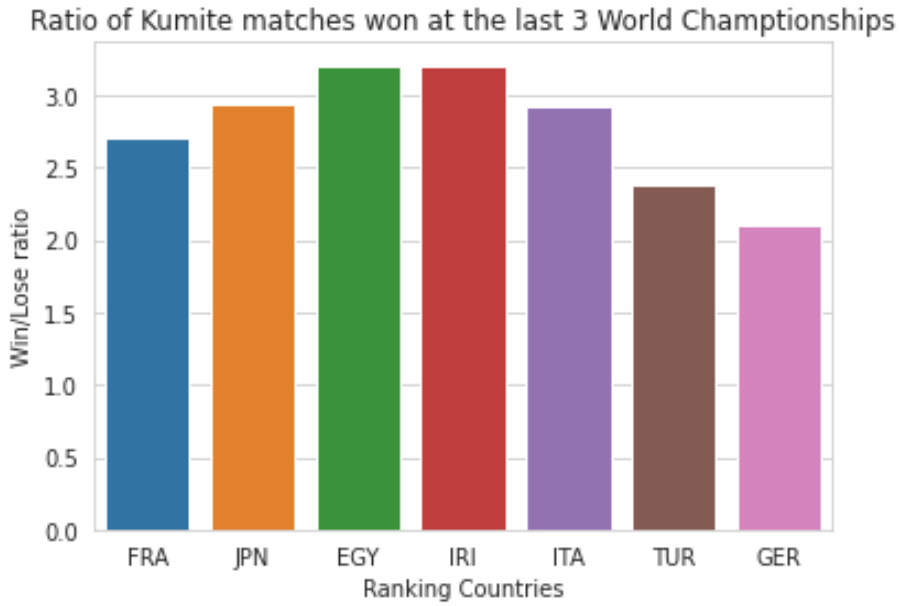

Figure 1: Countries with Kumite match win average >= 2 and >=1 1st place finish

The perceived top tier countries of Japan, France, Turkey, Iran and Egypt are included in this list. However Azerbaijan does not rank among these countries as it has an average matches won of less then two due to a high number of entrants that only gain a "Participation" ranking as shown below in Figure 2.

	matches_won	(count, 11th Place)	(count, 13th Place)	(count, 1st Place)	(count, 2nd Place)	(count, 3rd Place)	(count, 5th Place)	(count, 7th Place)	(count, 9th Place)	(count, Participation)
ranking_country										
AZE	1.965517	0.0	0.0	2.0	1.0	3.0	0.0	1.0	0.0	22.0

Figure 2: Azerbaijan matches won & ranking counts

Participation ranking indicates a competitor who entered the tournament but did not progress far enough to gain a numeric rank.

Germany and Italy were not initially considered top tier countries but do have average matches won of greater than or equal to two with at least one 1st place.

A breakdown of finishing places gives more detail to the nature of the success of these countries as shown in Figure 3.

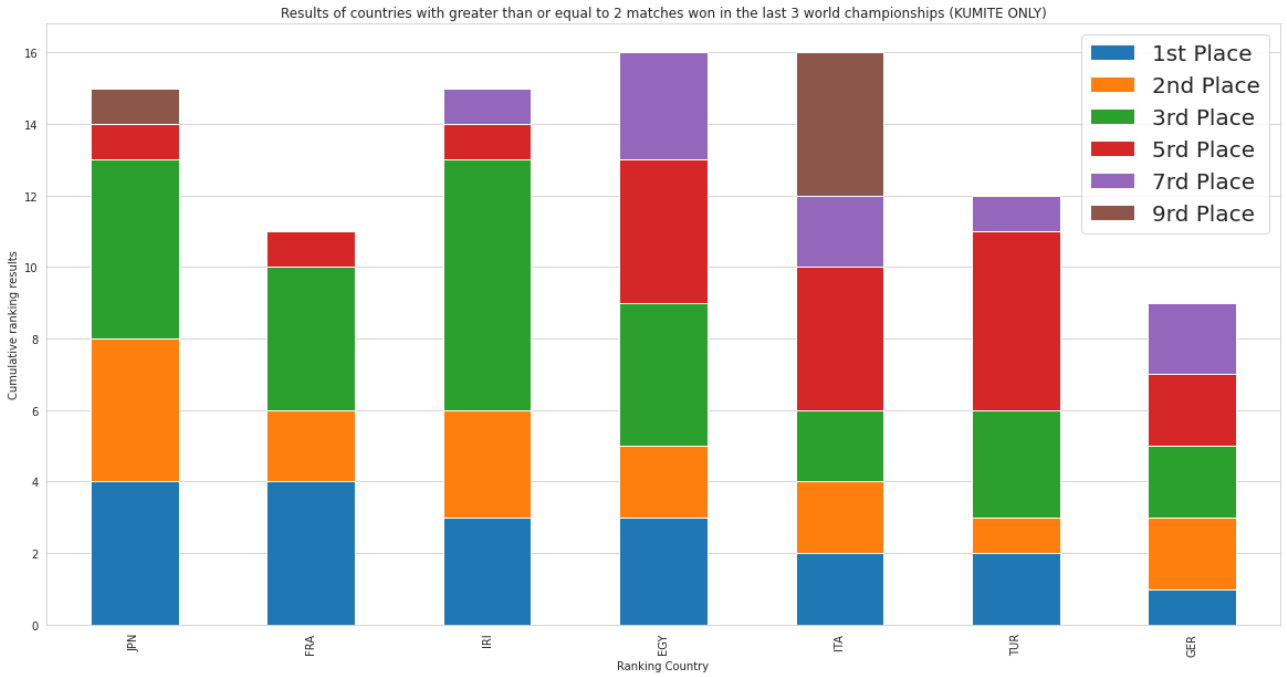

Figure 3: Cumulative ranking results of top tier countries

Germany is consistently able to win matches but fall short of winning many medals whereas Japan, France and Iran have high match win averages and consistent medal winning performances (1st, 2nd and 3rd place) . Egypt, Italy and Turkey are between these two groups with a mix of both.

Are the top tier countries specialists in a specific weight or sex?

The Kumite categories in WKF karate are separated by gender and split into a range of five weight categories for each. Given that the top tier countries have been identified in the previous question, these results can be from consistent performances in specific categories or success across a range of weights and both genders.

The Karate1 Premier League (WKF, 2021) is the most prestigious set of open competitions outside of the World and Continental Championships and mirror their categories. The medal winning performances (1st, 2nd and 3rd place) from of Karate1 tournaments from 1st January 2018 onwards were collected and split by gender and then weight.

Viewing the medal winning performances based on gender with identical scales, as seen in Figure 4, shows the most notable difference between genders for a country is Iran, with male medals approximately triple that of female medals.

Caution must be taken in assuming socio-economic reasons as the only factor for this disparity as the number of ratio of male to female entries is approximately 2:1. Conversely, Germany has an opposite male to female entrant ratio of 1:2 but this is not reflected in their results unlike Iran. France and Turkey have approximately 20% more male entrants with the remaining three countries having similar entrants.

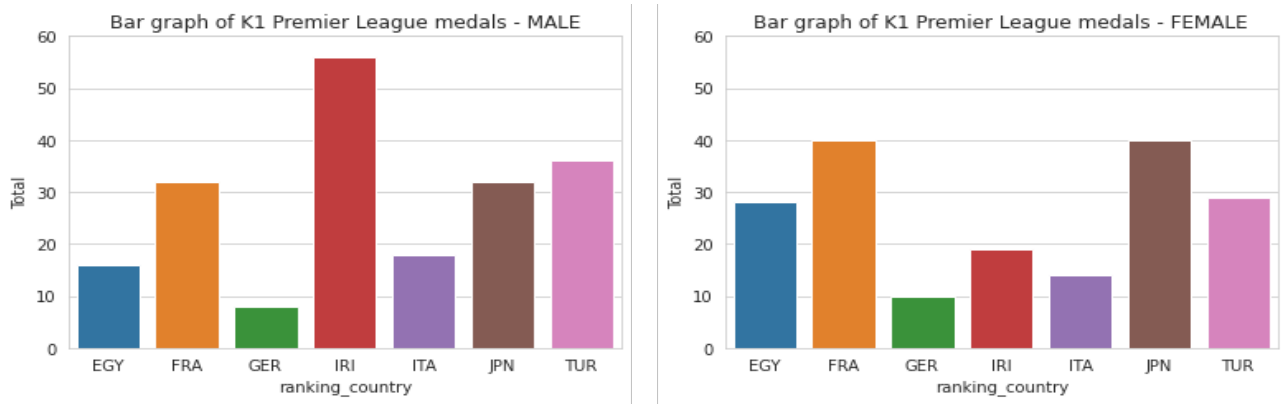

Figure 4: Gender comparison of Karate1 medals

Japan and Germany both produce marginally better results in female categories with Egypt producing marginally better results in male categories.

Both male and female categories have a range of five weights which have been redefined as Lightest, Light/Middle, Middle, Heavy/Middle and Heaviest. These weights include both genders so that a top tier counties training regimes and fight styles can be viewed as beneficial or detrimental within the five weights structure that the genders share. The medal performances for each of these weight categories is shown in Figure 5.

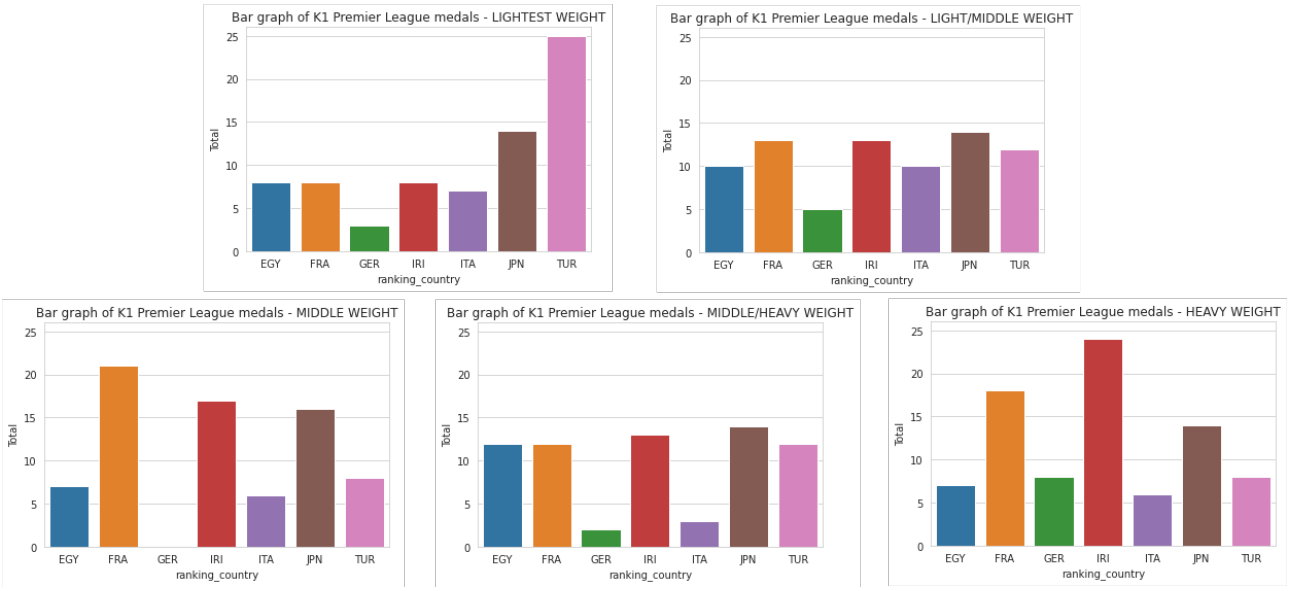

Figure 5: Weight categories comparison of Karate1 medals

Turkey particularly excels at the lightest weight categories whilst Iran and France have comparatively greater success at the heaviest weight categories. Germany has little or no success in the middle and middle/heavy categories which reflects why are consistently not as successful as the other top tier countries that. Japan never performs worse than third best across all weights.

These visualisations show that Iran is much specialised success with medium weight and heavier along with male categories producing the bulk of their medal success. Japan is the most successful all around countries with high medal distributions across both genders and all weight categories.

Does success in younger (Junior/Cadet/Under 21s) sections reliably lead to success in senior categories at the highest level?

Within the WKF competition structure there are tournaments and events aimed specifically at younger competitors. These tournaments help to develop competitors at the earliest age by allowing them to compete against an international pool of their own age group. Cadet events are restricted to 14 and 15 years old, Junior events are restricted to 16 and 17 years old and under 21s are restricted to 18, 19 and 20 years old.

Many competitors that entered the Senior World Championships entered these events as they developed at a younger age.

Using a K-NN (k-nearest neighbour) data mining algorithm with euclidean distance, the Junior/Cadet/U21 results of the last two senior world championships were analysed to understand if there are distinctive trends that would indicate future senior success from younger results.

Initially all the competitors who were entrants in the most recent two WKF Senior World Championships were collected and a status of “Medalled” or “Not-Medalled” applied based on if they finished 1st place, 2nd place and 3rd place or did not.

All the events for those competitors that had an event factor of two or more, were held not more than four year before the first WKF Senior World Championships to and up to the date of the most recent championships and were specifically either junior, cadet or U21 events were collected. All the finishing ranks from these events were then counted and combined with if they had medalled or not. The shape of this data is shown in Figure 6.

	ranking_competitor	1st Place	2nd Place	3rd Place	5th Place	7th Place	9th Place	11th Place	13th Place	Participation	Medalled?
0	ALB155	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	3.0	No
1	ALB2001	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	No
2	ALB2002	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	No
3	ALG194	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	No
4	ALG2016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	No
...	...	...	...	...	...	...	...	...	...	...	...
485	VEN265	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	5.0	No
486	VIE2003	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	No
487	VIE2030	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	No
488	VIE2069	1.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	1.0	No
489	WAL2014	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	No

Figure 6: K-NN classifier data-shape

The data on “1st place” finishes through to “Participation” was used as measures to identify any unique features that would indicate a ranking competitor to have medalled at a world championships. A leave-one-out algorithm was used were a single competitors “Medalled?” data was removed and all but one one of the competitors placement data was used to and identify similarities with known data by matching it with the data that it resembled the most.

It is known that the dataset contains a total of 490 competitors, 456 not having medalled at the world championships and 34 that have.

Using both normalised and normalised data the most successful version of the classification algorithm version was able to correctly classify 3 medalled competitors, a success rate of only 8.82%. All other versions either had fewer or no correct medalled classifications.

This extremely low success rate is for classifying successful medallists at world championships is due to the data from cadet, junior and U21 competitions not providing distinct features that can be found using K-NN data mining.

Therefore, success at junior, cadet or under 21 events does not reliably lead success at WKF Senior World Championships.

Does frequency of attendance of small tournaments correlate with world or continental federation success?

Selecting which events to attend is a decision that competitors frequently face with time, travel, tournament quality and funding constraints being factors in choice. Some competitors may attend every tournament possible while others are more selective.

A medal tally for each competitor and each country that have medals at either continental championships or world championships from 1st January 2012 onwards. A count for the number of lower events entered was also created for each competitor and country by counting all event entries from 1st January 2011 onwards including continental championships, world championships or invitational championships as some competitors have only attended these according to the rankings database.

Plotting the medals won against all competitions entered for each competitor with a regression line, as shown in Figure 7, indicates there is a positive relationship between the two variables. As one rises, so does the other. However, due to how densely packed areas of the graph are and an extremely high number of outliers away from the regression line it is difficult to tell just how strong the relationship is.

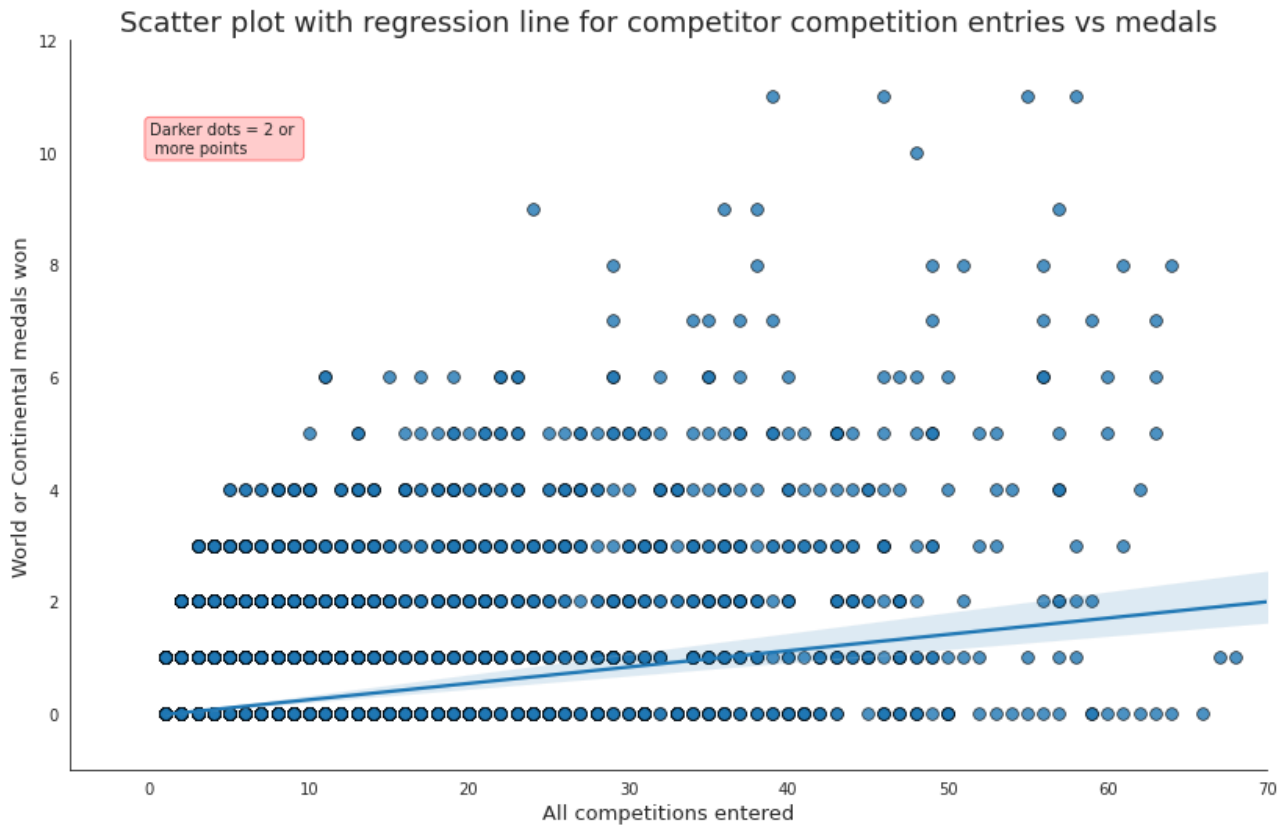

Figure 7: Elite competitor medals vs all competitions entered

Using Pearson's correlation coefficient the r number is calculated as 0.525 to 3 significant figures with a p number of 0. This means we can reject the null hypothesis that the number of competitions entered and the number of world or continental championship medals won

is unrelated and the trend is random. There is a medium to strong linear correlation between the two variables meaning that an increasing number of WKF ranking competitions entered will increase the number of medals won as a competitor.

This trend is mimicked when we view the data on a per country basis rather than individual competitors, as seen in Figure 8.

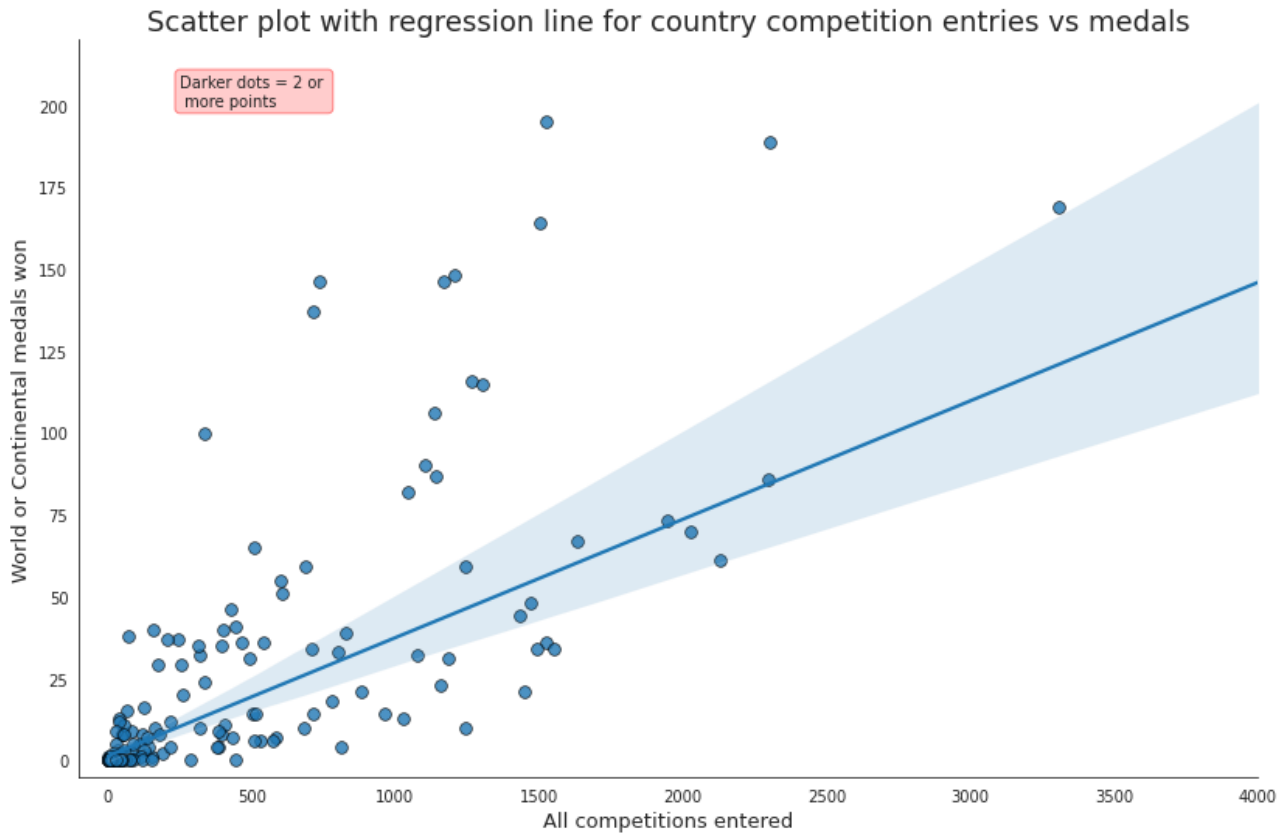

Figure 8: Elite country medals vs all competitions entered

Again, using Pearson's correlation coefficient the r number is calculated as 0.727 to 3 significant figures with a p number of $1.337e^{-29}$. This indicates an even stronger positive relationship between the number of competitions entered as a country and the number of world or continental medals won by a country.

Conclusion

The WKF rankings database is primarily designed as an online record of current WKF standings and as such conclusive outcomes from its analysis are limited. Due to the overwhelming and diverse range of competitors and events held within it the most useful analysis has come from meta based analysis which looks at general trends across the whole dataset. Attempts at more specific outcomes such as seeking trends in past event rankings to indicate future performances failed due to a lack of unique data points.