Contents

1 Abstract

Crossfit is a fairly new sport that became really popular in the last 30 years. It is a combination of various types of physical activity such as:

- Moving heavy weights
- Running for short and long distances
- Performing classical weightlifting movements like snatches
- Hand walking and other acrobatic movements

This wide range of challenges requires a wide range of physical skills. Ever since Crossfit was a thing, there has been a debate of which physical components are more important. We will try and help answering this question.

The various data components are analyzed to understand their distribution and inter correlations. We use statistical tests such as Kolmogorov smirnoff to check for normality. We also examine how the age of the athlete effects the performance. For that we use the Mann-Whitney U test.

Finally, we move to the most interesting question, what are the most important components to predict crossfit success. We pick a benchmark workout called Fran and build a linear regressing using the other physical skills to see which one will have the most influence on the final result.

Our results indicate that even though Fran is a high intensity endurance workout with relatively light weights, the strength physical skill has the highest influence on the final result.

2 Introduction

Crossfit has been founded by Greg Glassman in 2000 [1]. Ever since it became one of the biggest multinational fitness trends. There are around 10000 Crossfit gyms across 150 countries.

Unlike many other sports, crossfit (as it's name suggests) consists of a lot of different physical activity elements. This makes training for crossfit an optimization problem, every athlete has to decide how much time he will he spend working or raw power, endurance or acrobatic skills.

The data set [2] includes 423006 athletes with 27 features per athlete. We will focus on a subset:

- deadlift: The maximal weight the athlete can pull with the deadlift movement
- run5k: The fastest the athlete can run 5 kilometers
- age: The age of the athlete
- howlong: For how long has the athlete been training

- fran: The fastest the athlete can complete the Crossfit Fran workout

A few words about Fran, the objective of this workout is to complete as fast as possible:

- 21 thrusters (a barbell movement)
- 21 pullups
- 15 thrusters
- 15 pullups
- 6 thrusters
- 6 pullups

This workout combines power and a lot of endurance and we will use it as an indication of how good the athlete is in Crossfit. We will try to understand how important are the other features to get a good Fran score.

3 Results

3.1 Data cleaning

A brief examination of the data reveals there both garbage data and some unwanted outliers. For example here is the deadlift data before and after the cleaning. Prior to the cleaning the maximal deadlift result was 8388607 KG, way above the world record

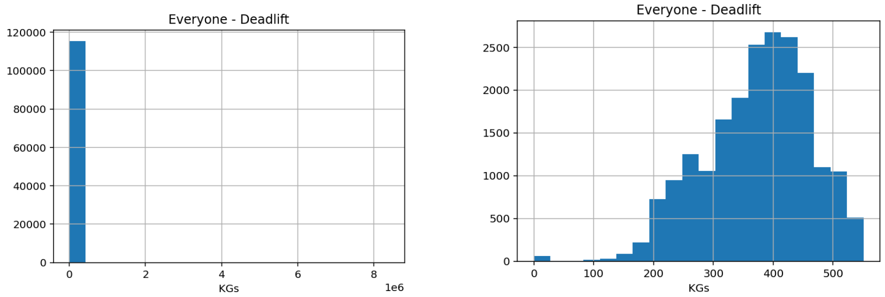


Figure 1: Original Deadlift distribution vs the distribution after the cleanup

Also, I wanted to focus this work on ”enthusiastic” athletes, a group that would hopefully include people who are not professionals but are also really passionate about the sport. To achieve this goal I removed the top and button 1% of the data for every column of interest.

Another important aspect is to split the woman and man athletes as this has a very high effect on the results.

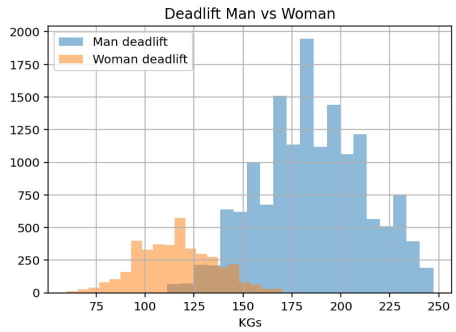


Figure 2: Man vs Woman deadlift distribution

3.2 Data distribution

First we examine the distribution of the various columns to see if we can assume normality in some cases.

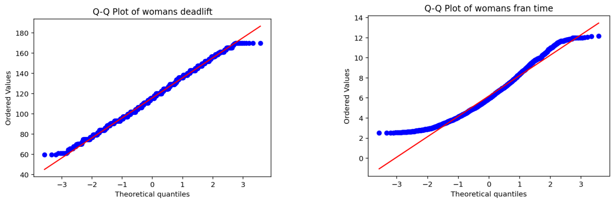


Figure 3: QQ plot of woman's atheletic data

I found it really interesting that the deadlift results (especially for woman seem to behave a lot like a normal distribution). To validate that point lets use several normality tests:

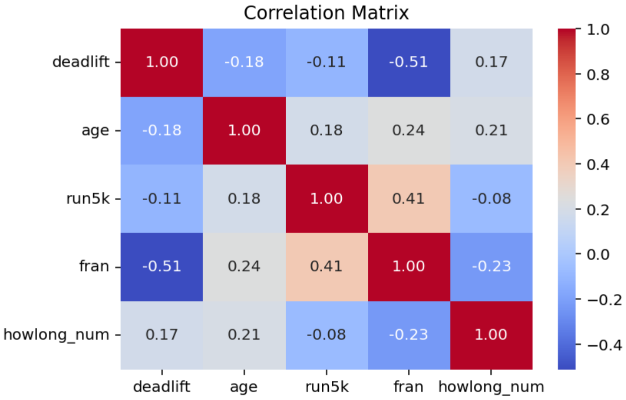
Table 1: Womans deadlift normality Test Results

Test	**Statistic**	**p-value**
Kolmogorov-Smirnov	0.0435	$p = 5.46 \times 10^{-7}$
D'Agostino and Pearson	25.8483	$p = 2.44 \times 10^{-6}$

Note that in these tests the null hypothesis is that the distribution is normal. So this low p-value indicates that the distribution in NOT normal. This illustrates how qq-plots might be misleading.

3.3 Data correlation

For this part lets focus on the man athletes. Lets examine the data correlation. Interestingly enough, the strongest correlation of every feature is with the Fran time. This represents in a way the fact that a crossfit workout is dependent on multiple factors. Those factor may be really different, like running 5K and deadlifting (-0.11 correlation), however, both of them have a strong influence on the Crossfit workout.



3.4 Age and its effects

At this stage I wanted to check how much effect did the age have on the Fran workout. In this analysis we will focus on the man's group. We will start by observing the change as the age progresses. Note the small lines of every bin represent the variance in that bin.

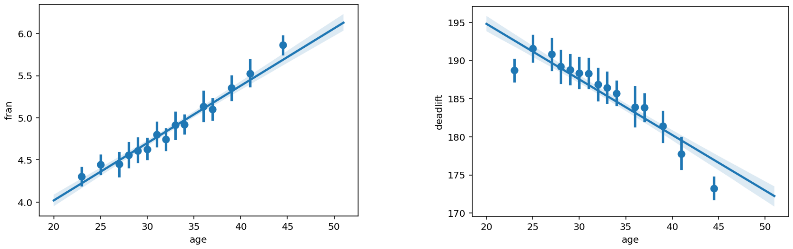


Figure 4: Performance degradation as a function of age

There is however something encouraging, it seems that the stronger you, the better your Fran result is

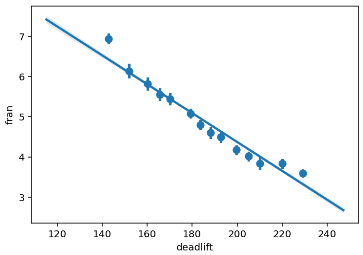


Figure 5: Fran as a function of deadlift

And indeed when comparing Fran for 2 groups:

$$O = \{x \in \text{Athletes} \,|\, x_{\text{age}} \geq 35 \ \wedge \ x_{\text{training}} \geq 2 \text{ years}\}$$
$$Y = \{x \in \text{Athletes} \,|\, x_{\text{age}} < 35 \ \wedge \ x_{\text{training}} \geq 2 \text{ years}\}$$

And a Mann–Whitney U test non parameteric test (as the distribution doesn't seem normal or any other distribution I know) agrees with this.

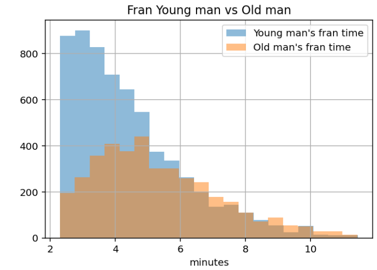


Figure 6: Fran as a function of deadlift

$$H_0 : \mu_{\text{young}} = \mu_{\text{old}}$$
$$H_1 : \mu_{\text{young}} < \mu_{\text{old}}$$

and the results are:

$$statistic : 8766562.5$$
$$p - value : 4.18e - 138$$

Meaning we can override H_0. However, if I change the groups to be:

$$O = \{x \in \text{Athletes} \,|\, x_{\text{age}} \geq 35 \ \wedge \ x_{\text{training}} \geq 2 \text{ years} \ \wedge \ x_{\text{deadlift}} \geq 180\}$$
$$Y = \{x \in \text{Athletes} \,|\, x_{\text{age}} < 35 \ \wedge \ x_{\text{training}} \geq 2 \text{ years} \ \wedge \ x_{\text{deadlift}} < 180\}$$

The results we get are opposite with Mann–Whitney U test proving this is significant:

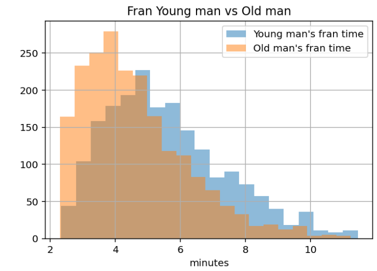


Figure 7: Strong old athletes vs Young weak athlete

3.5 Linear regression

The next part of the research is to try and understand the how much each feature impacts the Fran result. After examining the relationships they seem more or less linear:

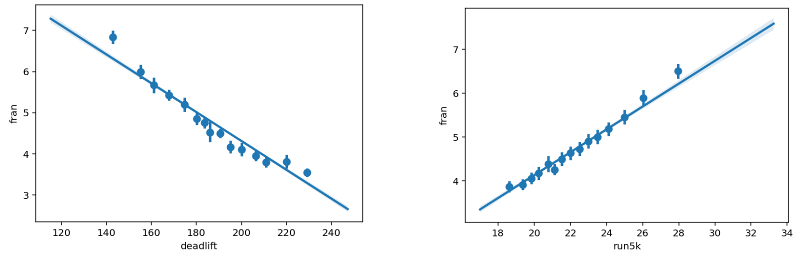


Figure 8: Fran as a function of the features

The subset of data we will work with, will be:

$$A = \{x \in \text{Athletes} \,|\, x_{\text{training}} \geq 1 \text{ years}\}$$

$$F = \alpha Age + \beta Deadlift + \gamma Run5K + \delta TrainedFor + C \tag{1}$$

Table 2: OLS Regression Results

Variable	**Coef.**	**Std. Err.**	**t**	**P>|t|**	**[0.025, 0.975]**
Intercept	4.7765	0.014	346.198	0.000	[4.749, 4.804]
Age	0.2755	0.015	18.753	0.000	[0.247, 0.304]
Deadlift	-0.7666	0.014	-53.284	0.000	[-0.795, -0.738]
Run5K	0.6136	0.014	43.295	0.000	[0.586, 0.641]
TrainedFor	-0.2522	0.014	-17.546	0.000	[-0.280, -0.224]

with an Adjusted R^2 score of 41%

4 Methods

4.1 QQ plot

It is a visual tool that is used test wether points behave according to a certain distribution. The idea is to draw them on a graph where the x axis are the quantiles and the y axis is how many data points are in that quantile. In this paper we've used this technique to check a certain data column for normality. Where if the points are on the $y = x$ line then the data behaves according to that distribution. It is a visual test so it is not precise.

4.2 Normality tests

Normality tests are used to check if the data in question comes from normal distribution. There are several such tests, each with it's pros and cons.

Name	**Description**	**Cons**
Kolmogorov-Smirnov	Compares the cumulative distribution of the theoretical distribution and the actual data. Uses a statistic $\sup \|\text{DataCumulative}(x) - \text{Theoretical}(x)\|$, which measures the largest distance between the two functions.	Less sensitive to differences in the tail.
D'Agostino and Pearson	Uses the skewness and kurtosis to determine if the data is normally distributed.	Some non-normal distributions might have the same skewness and kurtosis [3].

Table 3: Normality Tests and Their Characteristics

4.3 Correlation matrix

Tests all the pairs of features and display how correlated each pair is. There are several methods to measure correlation. In this paper pearson's correlation was used. It is defined as:

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

4.4 Mann-Whitney U test

This is a non parameteric test (meaning it doesn't require the assumption that the data distributes in a certain manner). The objective of this tests is to determine whether group A and group B of independent data have the same of different mean. While it is also possible to specify the type of difference ($A > B$ or $B > A$). It is done by:

- Give a rank to all the data points (smallest is 1 etc..)

- Compute $U_1 = n_1n_2 + \dfrac{n_1(n_1+1)}{2} - R_1, U_2 = n_1n_2 + \dfrac{n_2(n_2+1)}{2} - R_2$. Where $R_{1,2}$ are The sums of the ranks of each group.
- Next we take as the statistic the minimal value between $U_{1,2}$. The distribution of this value is known under the Null Hypothesis (that assumes both A and B are of the same distribution)
- We calculate the p-value and determine whether H_0 should be rejected

4.5 Linear regression

This is a model that explains a certain variable Y by a set of other features $X_1, X_2, .., X_n$. It assumes they have a linear relationship which results in:

$$Y = \sum_{1}^{n} \alpha_i X_i + C$$

The model finds the best α_i and C for this kind of explanation. It is useful to normalize all the values of Y and X_i. Then the size of the corresponding α_i may be used as an indicator of it's significance.
Another term we've used is the AdjustedR^2. This is a very useful metric to understand how good your model is. R^2 measure the amount of variability that the model explained. The AdjustedR^2 takes into account the fact that R^2 will grow by simply adding additional features. It uses a simple normalization factor to address this.

5 Discussion

The research has basically addressed two questions:

- Age: the effects of age and how can they be overcome
- Physical components: which physical components are the most important ones when trying to excel in Crossfit.

For the first question it seems that age has an effect in almost every parameters. It seems to hurt the athlete in every possible category. However, it does seem like older athlete who are much stronger then average seem to get much better results then their younger and weaker counterparts. This illustrates that excelling in a certain physical component might compensate for being old.
For the second question we can observe that the order of significance is:

- Deadlift with 0.76, negative influence
- Run5K with 0.61, positive influence
- Age with 0.27, positive influence

- Trained for with 0.25, negative influence

This enforces our previous conclusion and gives hope that working out is more important then not aging.

6 References

References

[1] CrossFit, Wikipedia https://en.wikipedia.org/wiki/CrossFit.

[2] Kaggle, crossfit athletes https://www.kaggle.com/datasets/ulrikthygepedersen/crossfit-athletes.

[3] D'Agnostino and Pearson false positive https://stats.stackexchange.com/questions/62291/can-one-measure-the-degree-of-empirical-data-being-gaussian/62320#62320.