Abstract

PLACEHOLDER - use 250 words or less to summarize your problem, methodology, and major outcomes.

Using a dataset that contains over 200,000 observations representing powerlifters competiting at weight lifting competitions, our team attempted to create a regression model to accurately predict the total amount of weight lifted using only a handful of predictors available for each competitors. Our analysis revelaed that...

Key Words

PLACEHOLDER - select a few key words (up to five) related to your work.

- Kaggle
- Powerlifting
- Multiple Imputation
- Multiple Linear Regression
- 5

Introduction

A dataset from Kaggle (https://www.kaggle.com/open-powerlifting/powerlifting-database/home) contains observations on the peformance of competitors at various powerlifting events. The data contains 209,661 observations each one containing 5 predictors and 1 response variable. Our team's goal is to develop a regression that will accurately predict the final score of a powerlifter at a competition given that they were not disqualified. A powerlifter's score is in the form of kilograms, the sum of the total weight they lifted in the Squat, Deadlift and Bench portions of the competition.

The first 10 rows of the data is sampled below.

Federation	Sex	Equipment	Age	BodyweightKg	TotalKg
365Strong	F	Wraps	47	59.60	138.35
365Strong	F	Single-ply	42	58.51	401.42
365Strong	F	Single-ply	42	58.51	401.42
365Strong	F	Wraps	28	62.41	392.36
365Strong	F	Raw	60	67.31	383.28
365Strong	F	Raw	60	67.31	383.28
365Strong	F	Wraps	52	65.95	283.49
365Strong	F	Wraps	24	65.50	340.20
365Strong	F	Wraps	56	71.21	292.56
365Strong	F	Wraps	56	71.21	292.56

Literature Review

PLACEHOLDER - discuss how other researchers have addressed similar problems, what their achievements are, and what the advantage and drawbacks of each reviewed approach are. Explain how your investigation is similar or different to the state-of-the-art. Please do not discuss paper one at a time, instead, identify key characteristics of your topic, and discuss them in a whole. Please cite the relevant papers where appropriate.

Methodology

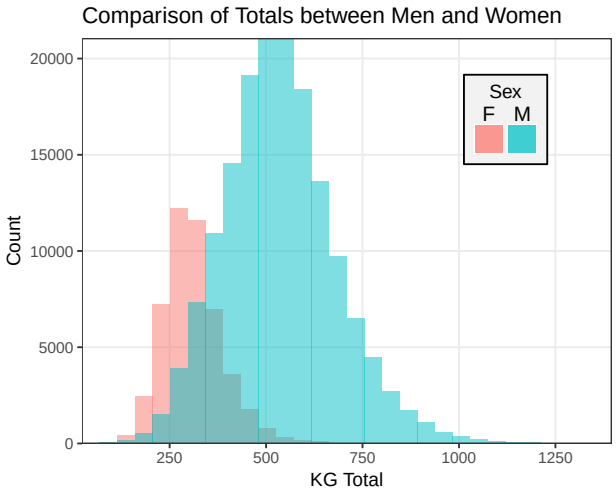
Data Exploration

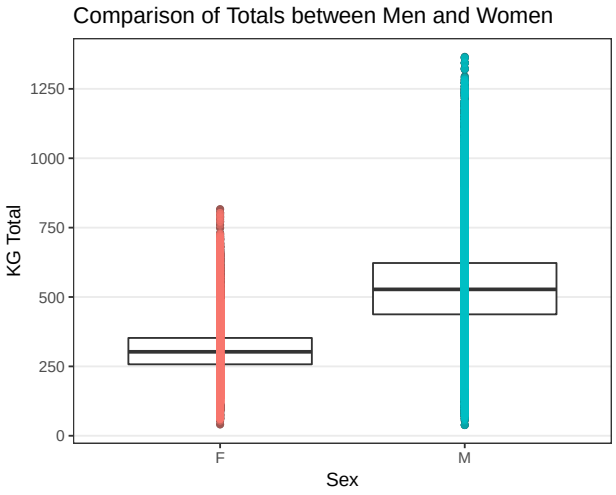
Sex

This predictor is categorical and is stored as either F for Female or M for male.

Appoximately 75% of the competitors are male and there is a clear seperation between the totals when comparing men and women. Both distributions are roughly normal and feature a number of upper and lower outliers.

Sex	n
F	48284
M	161377



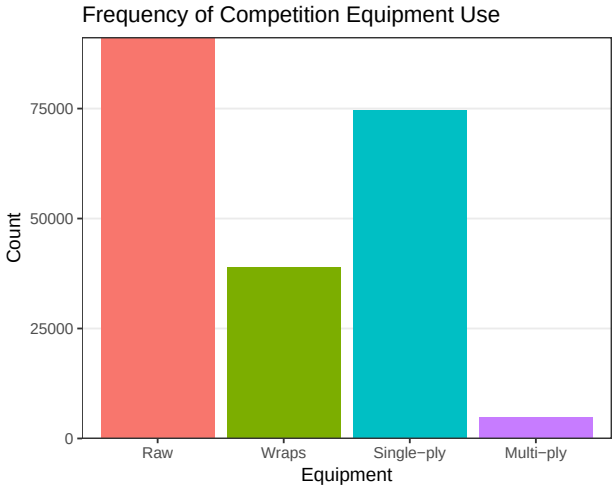


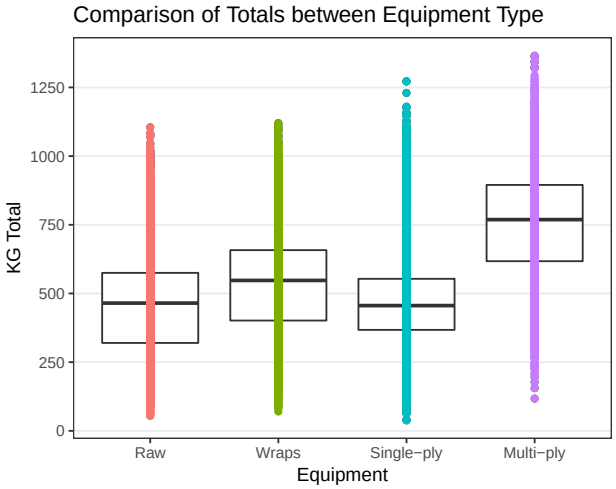
Equipment

There are four different lifting categories, broken down by the equipment that is allowed to be used during the competition. Different equipment can have a significant impact on the amount of weight a competitor can lift. The categories are:

- Raw: Only a lifting belt is allowed
- Straps: Wrist straps plus everything above is allowed
- Single-ply: A single-ply lifting suit plus everything above is allowed
- Multi-ply: A multi-ply lifting suit plus everything above is allowed

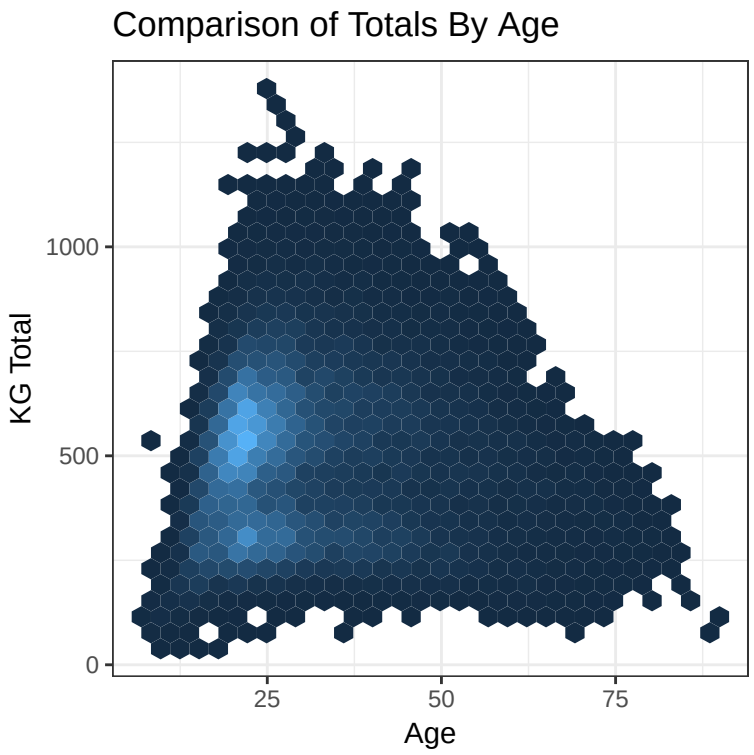
There is a clear separation in the quantity of competitors that use certain equipment, however there is not a noticable difference in the total weight lifted outside of the use of Multi-ply. It is unclear whether these predictors will be significant.





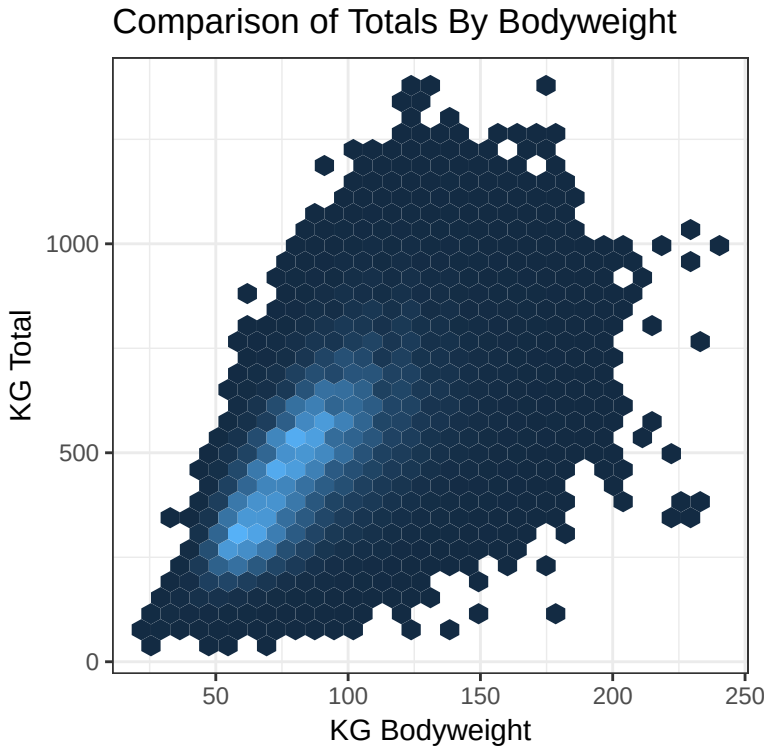
Age

The youngest competitor in the data set is 7 and the oldest is 90. It is likely that age will play a factor in total weight lifted as competitors in their prime years will be able to lift more weight. The below graph supports the idea that there is a growing and shrinking effect. That is, total weight lifted is fairly small and then grows as the competitors get older. Around the age of 50 however, natural declining of the body bring the total back down. It is possible that a second term may be needed for age account for this.



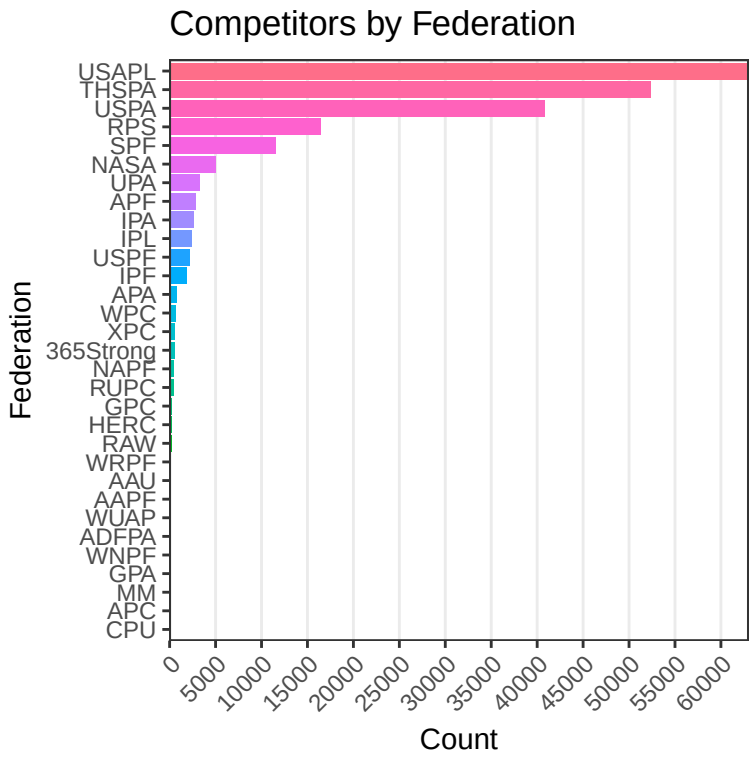
BodyweightKg

A competitor's weight should have a strong effect on their lifts. This is such a major factor that in powerlifting competitions, competitors are seperated into divisions based on their weight. Heavier competitors likely have more muscle and certainly have more mass. They are also likely to be taller and broader. However, bodyweight may be a problematic measurement. (See Below)



Federation

There are a number of federations that run competitions. A federation is an organization with it's own structure and rules for competition. This is an open data set and thus any federation, no matters it's size, can submit data. As can be seen in the below plot there are a few big federations and many more smaller ones. The rulesets between the federations are a mess and some of the smaller federations do a poor job of enforcing their rules. Furthermore, there are far too many federations to use them as a factor in the regression.



Weigh In Times

In order to try to capture the information from the federations despite their large numbers and discrepent and poorly enforced rule sets, we decided to use an unambiguous measure. Powerlifting competitions require competitors to weigh in to demonstrate their ability to lift in a selected division. When this weigh in occurs is critical. The earlier the weigh in is from the competition, the more time the competitor has to manipulate their weight afterwards to give them an advantage.

That is, a competitor weighing in at 150kg 2 hours before the competition and 24 hours before the competition will actually compete at far different weights.

Each federation either has weigh ins 2 hours or 24 hours before the competition. This information was collected via google searches and added to the data set. We anticipate that federations that allows earlier weigh ins should have higher totals as competitor's true lifting weight may be significantly higher than the reported weight.

Weigh	n
24h	13
2h	5

Steroid Testing

We can't perform an analysis on powerlifting without discussing the elephant in the room. Steroid use is rampent in powerlifting and the various federations have taken different stands on their use. Generally speaking federations are either 'tested' or 'untested'. However, the truth is much more complicated than that.

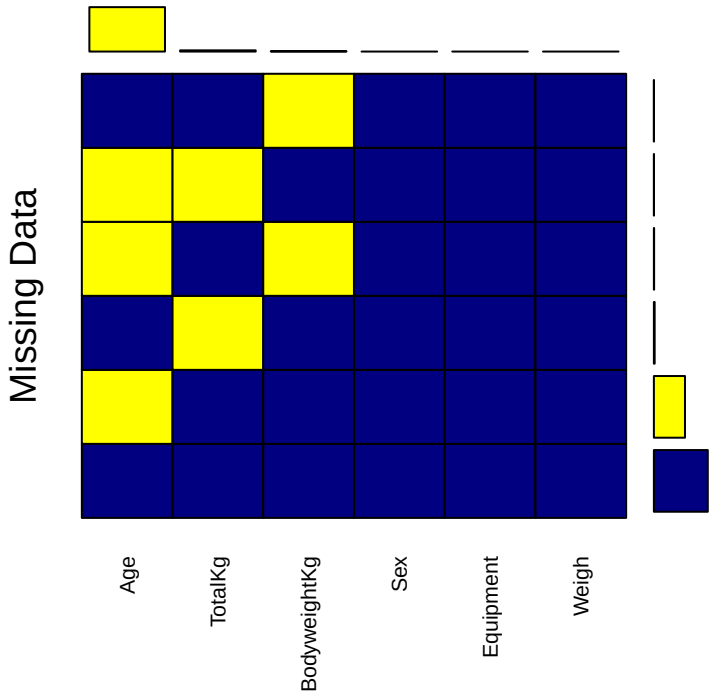
Different federations have different banned substance lists. Some federations have both 'tested' and 'untested' divisions. Some federations technically test but allow for large loopholes that allow competitors to avoid

detection. Some federations require 3 years of clean drug tests to compete but the lingering effects of steroids can have positive effects for upwards of 5 years. In short, it is not feasible to include a meaningful predictor on steroid use for our data set.

Data Preperation

Exploring the data resulted in us discovering that over 35% of the age data is missing in addition to a miniscule amount from the BodyweightKG predictor. Furthermore, there is a small proportion of reponse variable that is missing. We will seperate out this data and use it as our evaluation data set.

	x
TotalKg	0.0098361
Sex	0.0000000
Equipment	0.0000000
Age	0.3649178
BodyweightKg	0.0030242
Weigh	0.0000000




```
##
##  Variables sorted by number of missings:
##       Variable Count
##            Age 48749
##        TotalKg  1314
##   BodyweightKg   404
##            Sex     0
##      Equipment     0
##          Weigh     0
```

The Age and BodyweightKg predictors were imputed and the observations missing the response variable TotalKg were placed into an evaluation data set.

From there we created a training and testing set. The training partition will contain 80% of the observations while the testing partition will contain the remaining 20%.

With the data preperation complete we have three data sets. The training set contains 105822 observations, the testing set contains 26453 observations and the evaluation set contains 1314 observations.

Experimentation and Results

Model Building

Model Selection

Discussion and Conclusions

PLACEHOLDER - conclude your findings, limitations, and suggest areas for future work.

References

PLACEHOLDER - be sure to cite all references used in the report (APA format).

Appendix