Abstract

A dataset by Kaggle (here)[https://www.kaggle.com/open-powerlifting/powerlifting-database/home] contains observations on the peformance of athletes at various powerlifting competitions. 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 are 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

Data Exploration

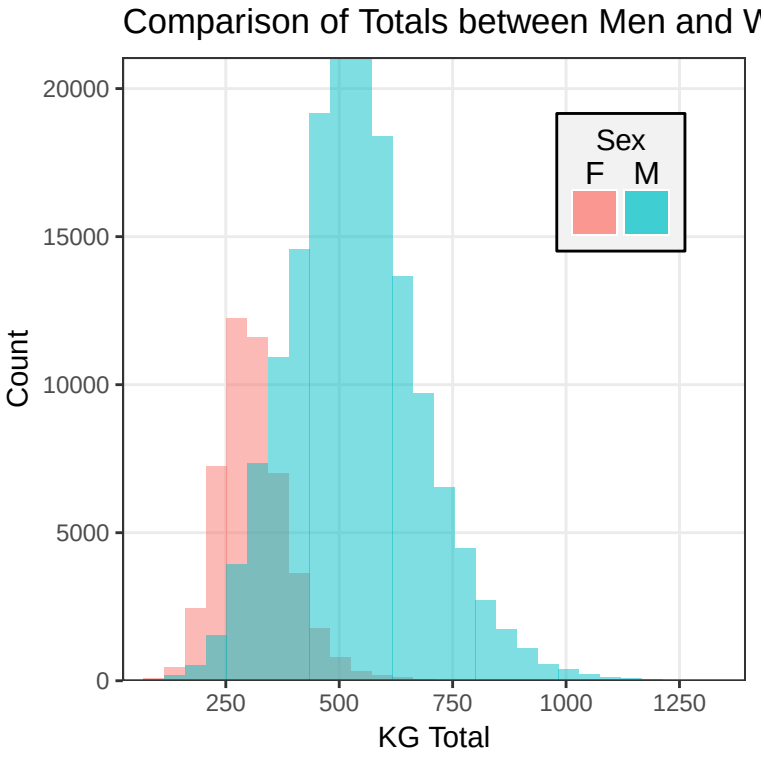
We began by performing a thorough analysis on each predictor.

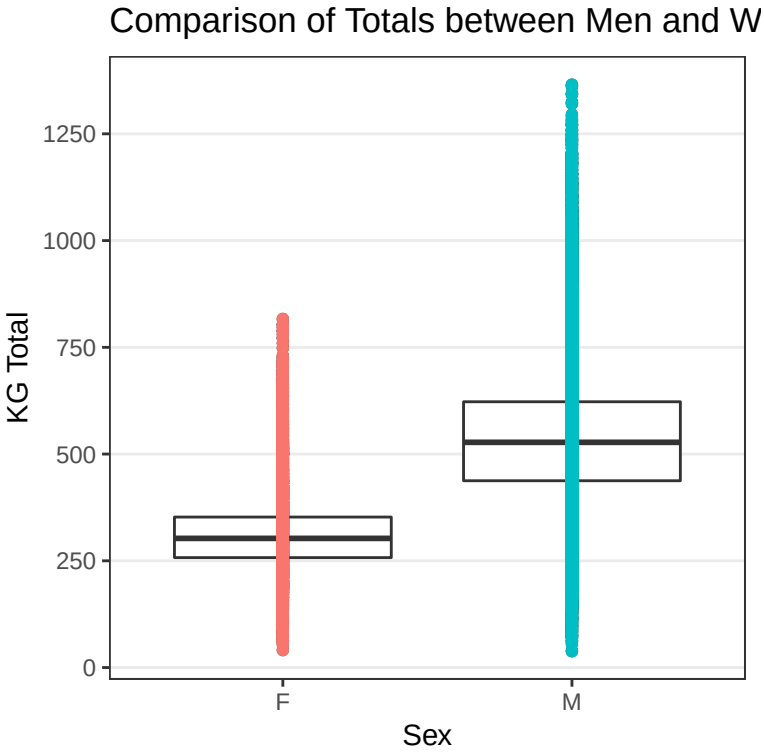
SEX

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

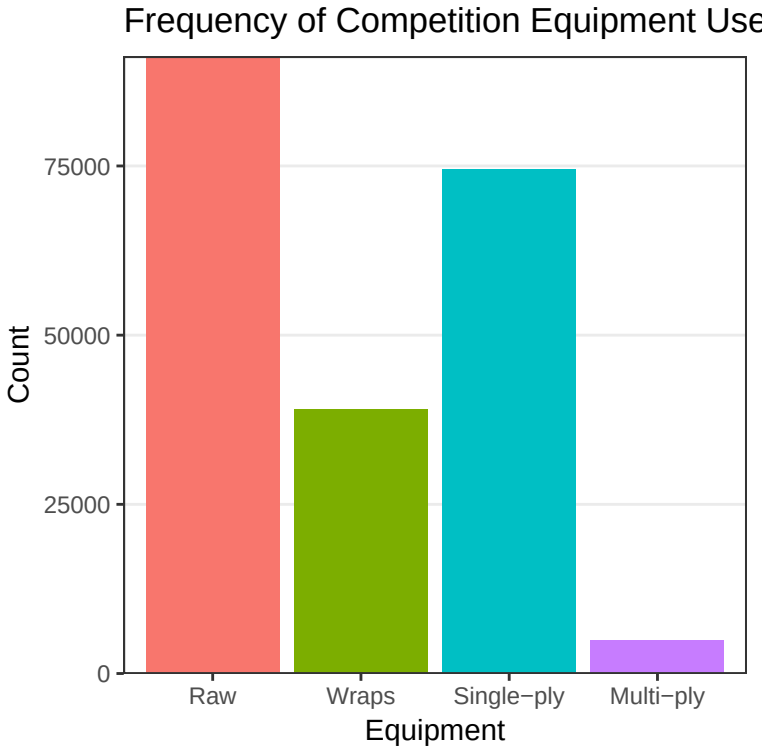

Comparison of Totals between Men and W
Sex
F
M
Count
20000
15000
10000
5000
0
250
500
750
1000
1250
KG Total

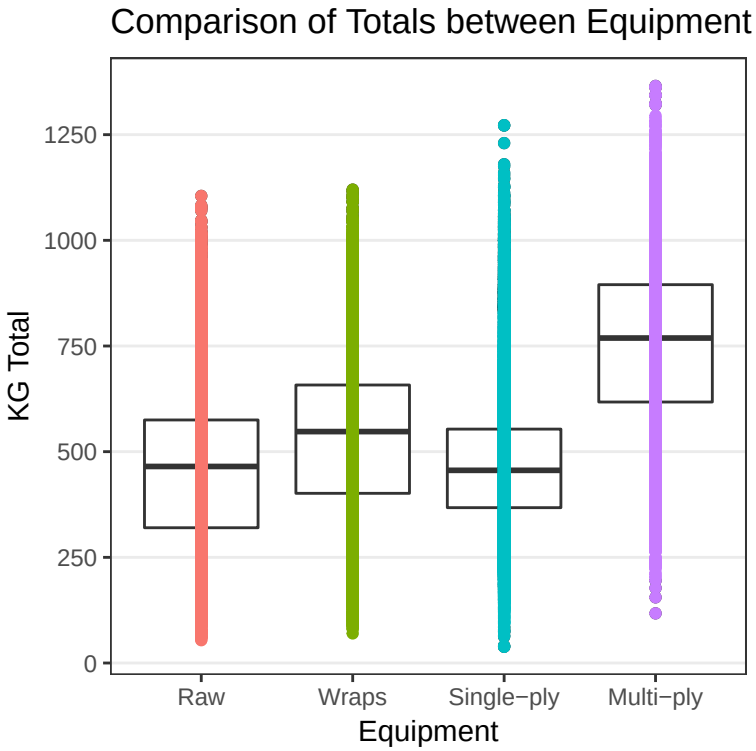

Comparison of Totals between Men and W
KG Total
1250
1000
750
500
250
0
F
M
Sex

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 marked difference on the amount of weight a competitor can lift. The categories are:

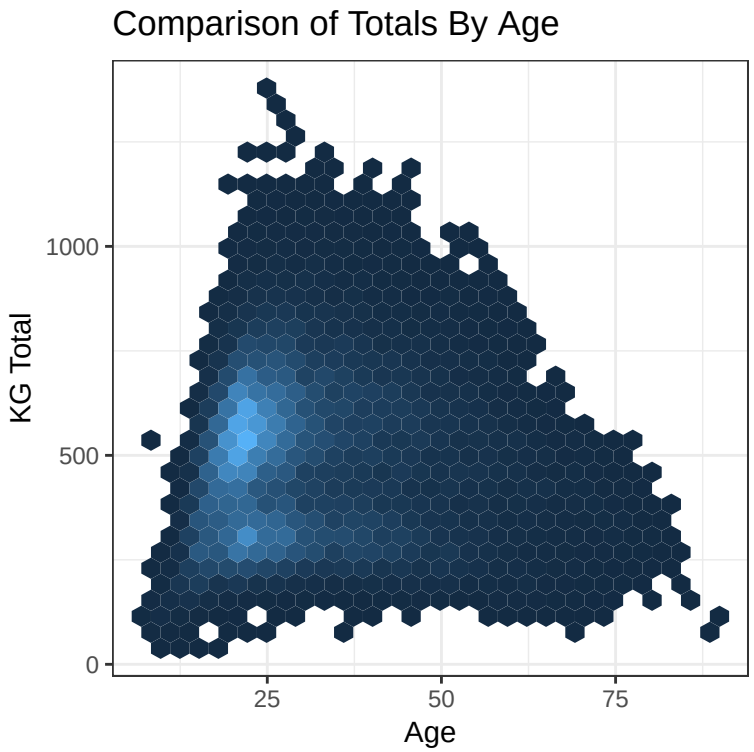
Raw: Only a belt is allowed Straps: Wrist straps plus everything above is allowed *Single-ply: A single-ply lifting suit plus everything above is allowed* Multi-ply: A multi-ply lifting suit plus everything above is allowed





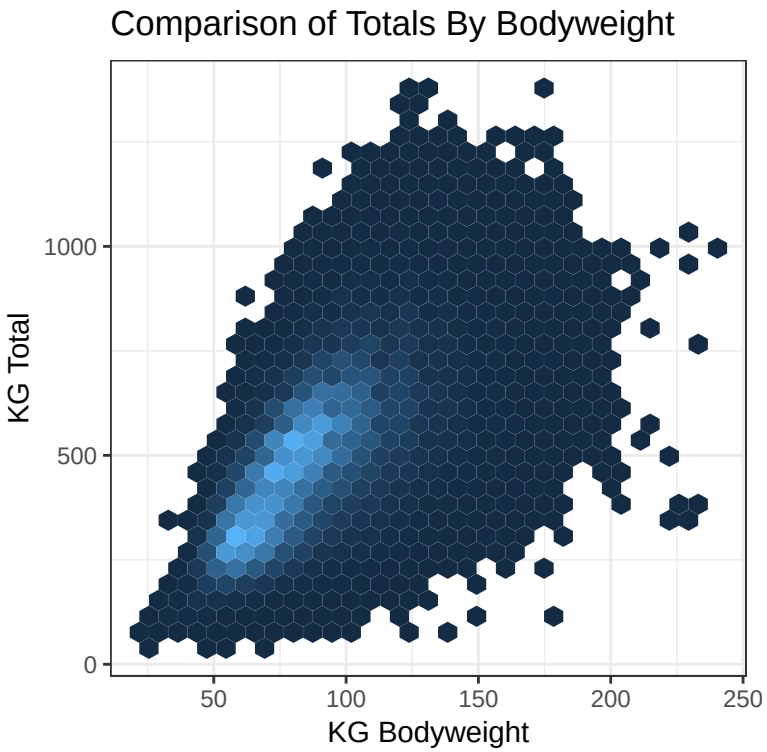
Age

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.



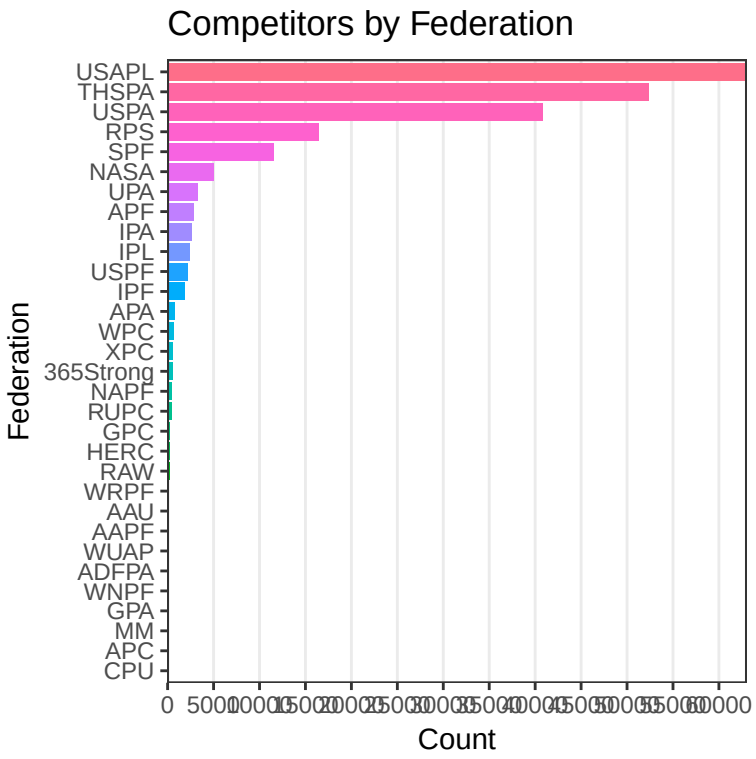
BodyweightKg

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 based on their weight. Heavier competitors likely have more muscle and certainly have more mass. They are also likely to be taller and broader.



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. A federation may result in a measurable different in total weight. Some federations are designed for beginners or offer a more family friendly atmostphere while others aim towards creating fiercer competition. Some federations are also more strict with drug testing than others.

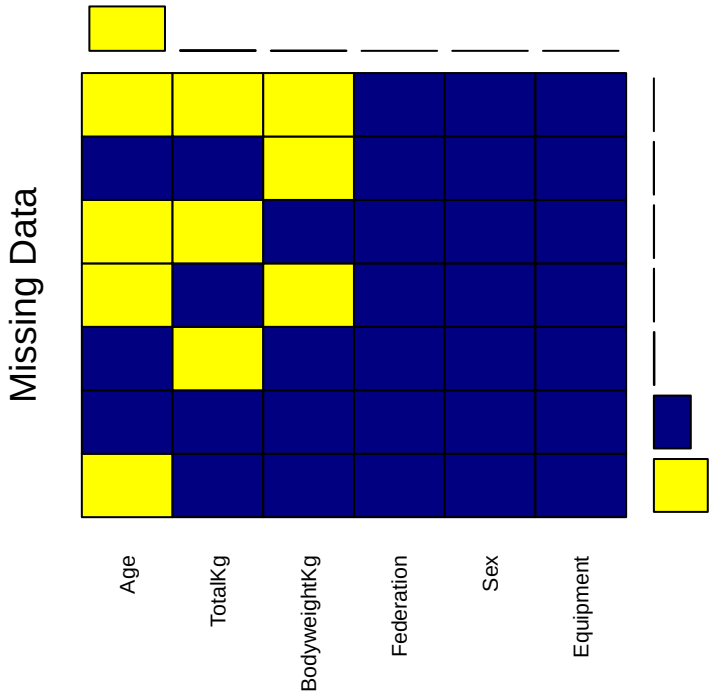


Data Preperation

Exploring the data resulted in us discovering that nearly 60% of the age data is missing. As useful as this data may have been, there is simply too much missing to consider imputation. Unfortunately, we will have to eliminate it from our data set. 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.

	x
Federation	0.0000000
Sex	0.0000000
Equipment	0.0000000
Age	0.5913785
BodyweightKg	0.0020748
TotalKg	0.0073738

```
## Warning in plot.aggr(res, ...): not enough horizontal space to display
## frequencies
```




```
##
##  Variables sorted by number of missings:
##       Variable  Count
##            Age 123989
##        TotalKg   1546
##   BodyweightKg    435
##     Federation      0
##            Sex      0
##      Equipment      0
```

Revisiting the Federations category shows that are still 23 federations represented in our data. This is simply too many to use the federation as a category. We will restrict the data to only include the two most representative federations, USPA and USAPL. Together they account for nearly 80% of the available data.

Federation	n
USAPL	62992
THSPA	52335
USPA	40889
RPS	16418
SPF	11565
NASA	5059
UPA	3280
APF	2823
IPA	2613
IPL	2415
USPF	2177
IPF	1930
APA	831
WPC	646

Federation	n
XPC	612
365Strong	542
NAPF	478
RUPC	473
GPC	293
HERC	288
RAW	214
WRPF	177
AAU	143
AAPF	123
WUAP	83
ADFPA	76
WNPF	75
GPA	46
MM	31
APC	28
CPU	6

This still leaves a small portion of missing BodyweightKg predictors. This data will be imputed. We will also split the data into a training set and a testing set for use in our regression modelling.

With the data preperation complete we have three data sets. The training set contains 52852 observations, the testing set contains 13211 observations and the evaluation set contains 1192 observations.

Model Building