**Describe your data source.** I got my dataset from Kaggle.com, a website that is a crowd-sourced platform for various datasets and online community for data scientists. It was founded in 2010 and it continues to host data science competitions today.

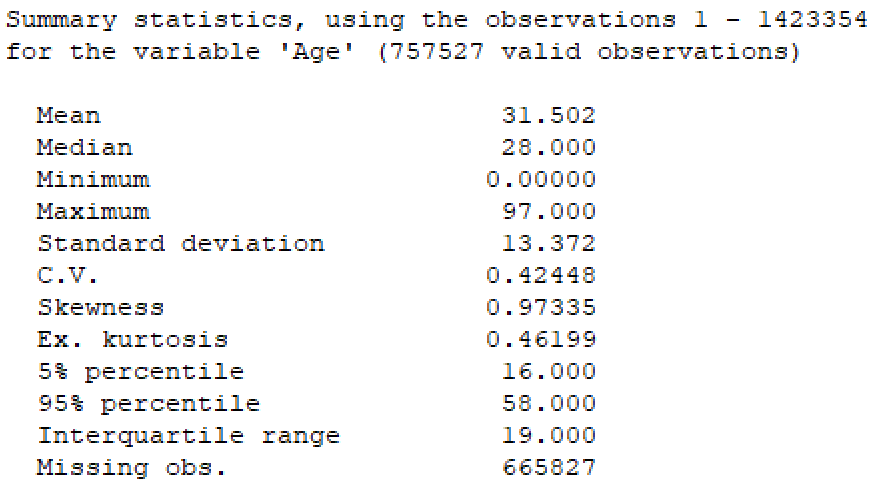
**When were the data collected?** The data in my set was collected from weight lifting competitions over the last 40 years. I’m thinking about cutting out any data from earlier than 2000 because how we train to lift has changed since the 70s-90s.

**What is the unit of observation (i.e. element/entity**)? Sex (m/f), age (years), bodyweight (kg), best of 3 squat (kg), best of 3 bench (kg), best of 3 deadlift (kg)

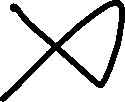
**For what purposes were the data collected?** The website is an online record of weight lifting competitions and meets. Individual competitors can look at where they rank among their peers and submit proof for adjustments as needed. They also keep a record of what published projects used their data. [Powerlifting Rankings (openpowerlifting.org)](https://www.openpowerlifting.org/)

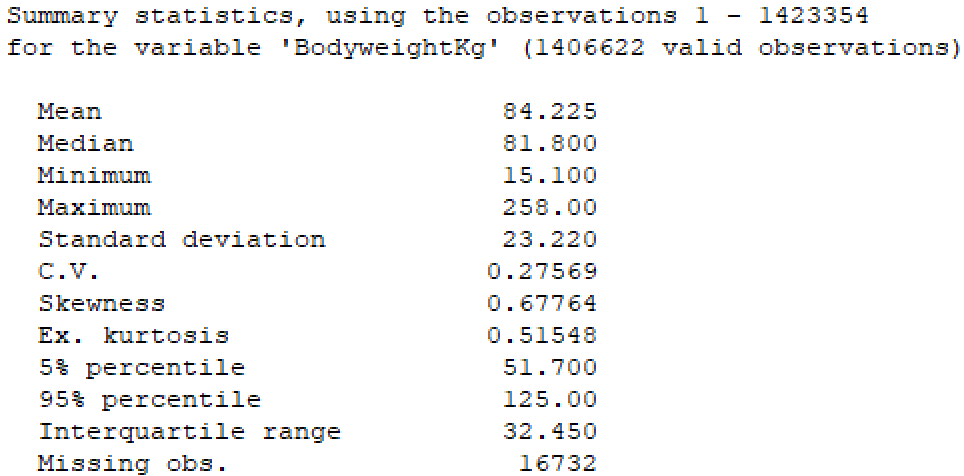
**Are there any well-known general limitations/issues with the dataset?** The two things that stick out to me is that: (1) If a competition doesn’t differentiate between certain things (using wraps vs not) then it’ll default to the lowest common denominator because the competition officials will not record/report the information. This can lead to lifts not being represented accurately. (2) The data in this set are representative of professional weightlifters. In other words, there is a lack of data regarding average or amateur weightlifters.

Original Dataset Amended Dataset (only March-April 2019, ages 20-50);

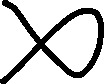
Text

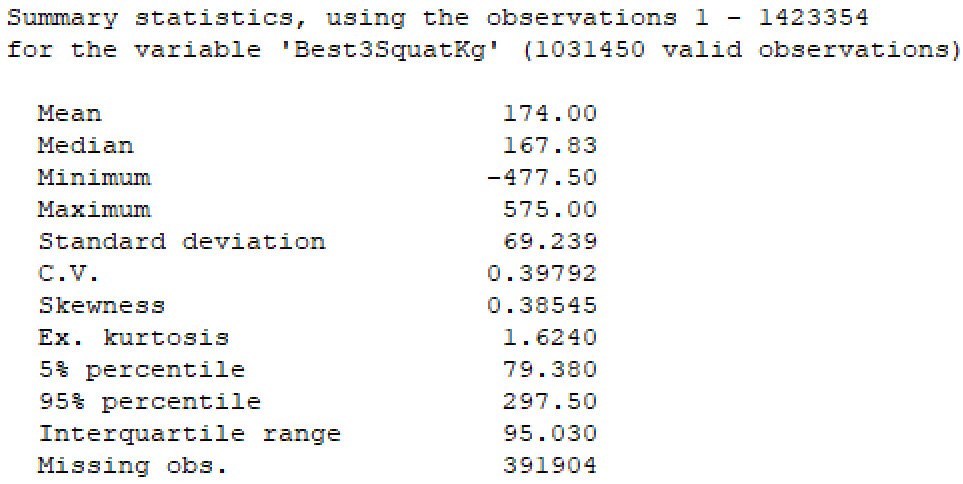
Description automatically generated



Text, table

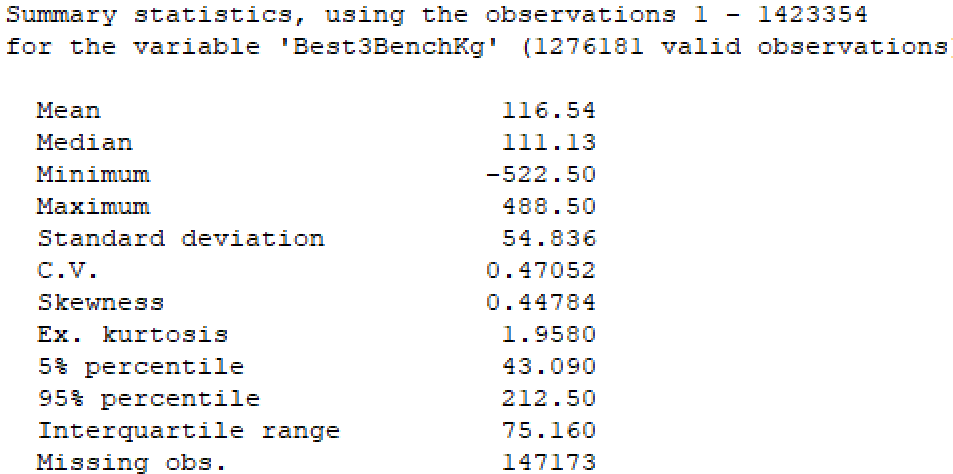
Description automatically generated



Text, table

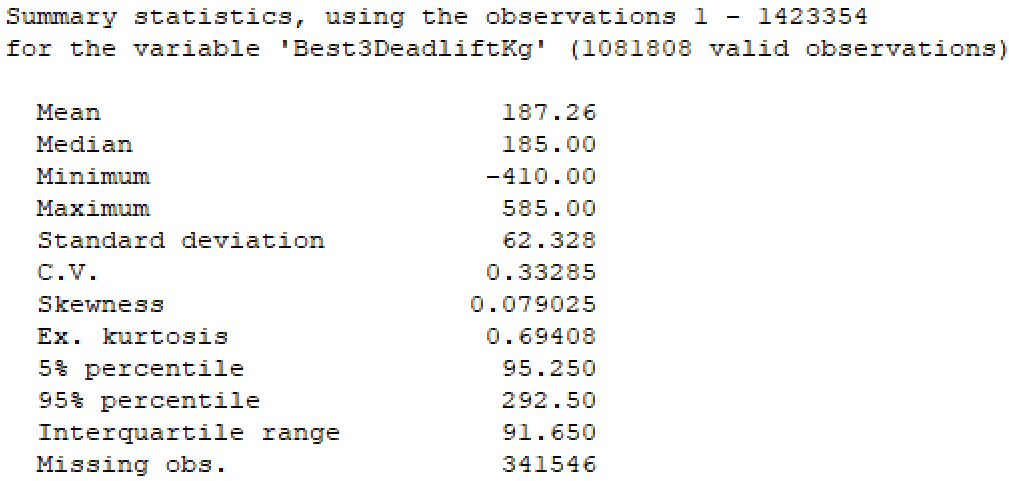
Description automatically generated with medium confidence



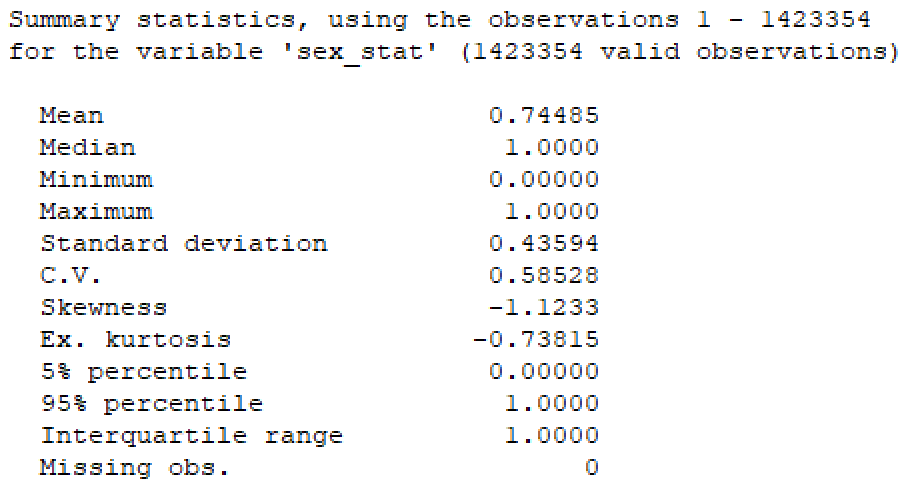
Table

Description automatically generated

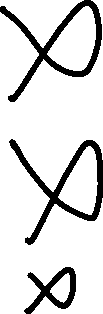


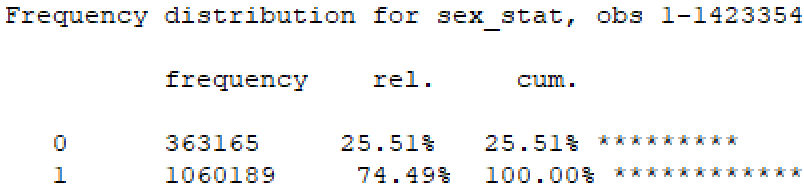
Text, table

Description automatically generated with medium confidence

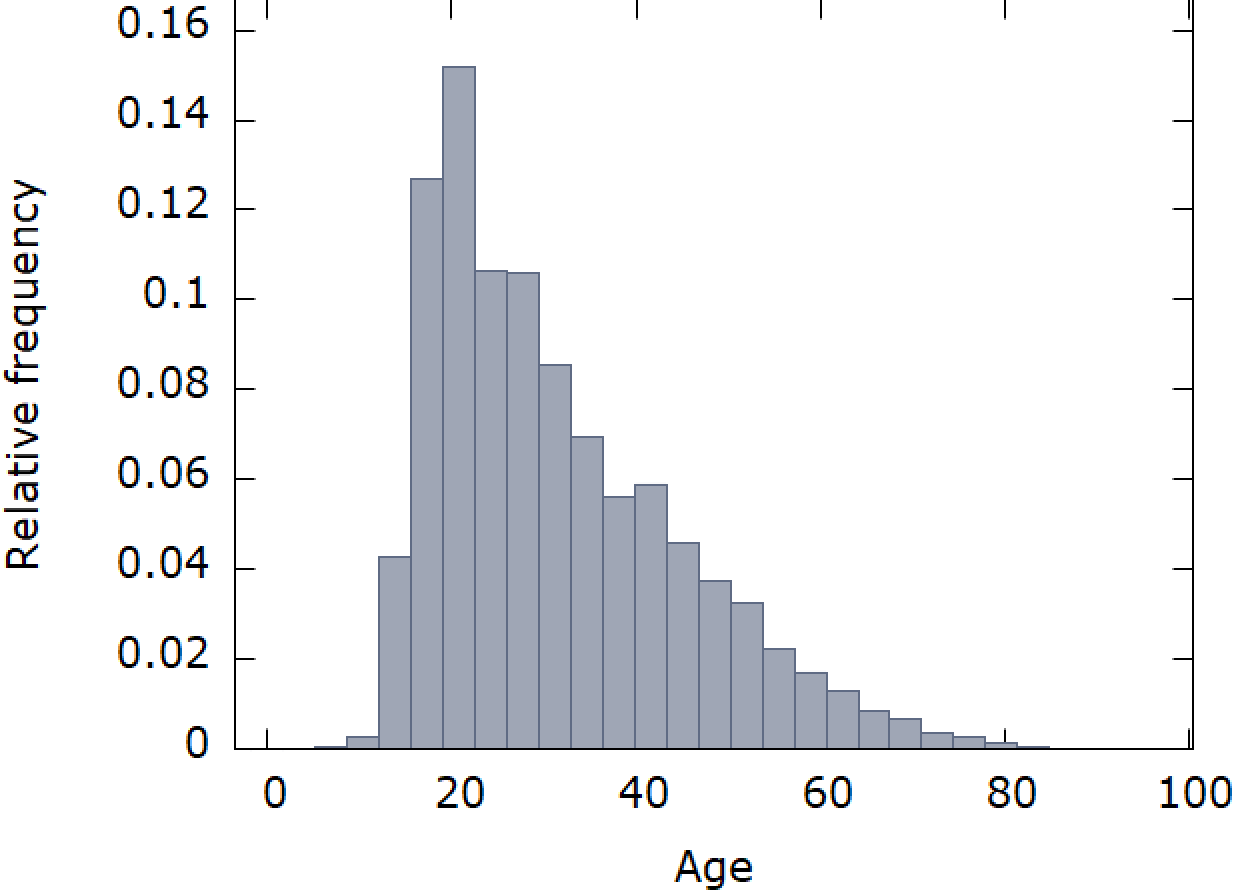
 Graphical user interface, text

Description automatically generated

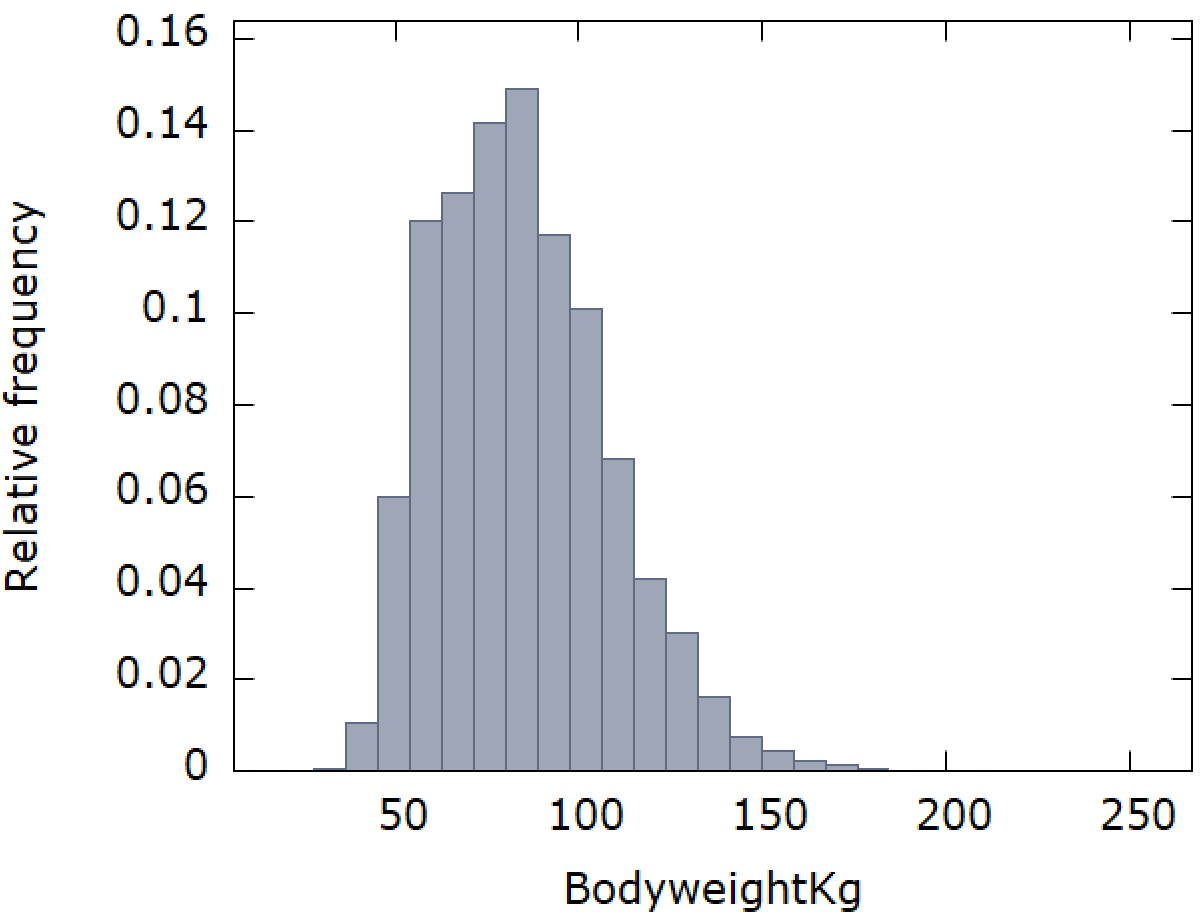


 Text

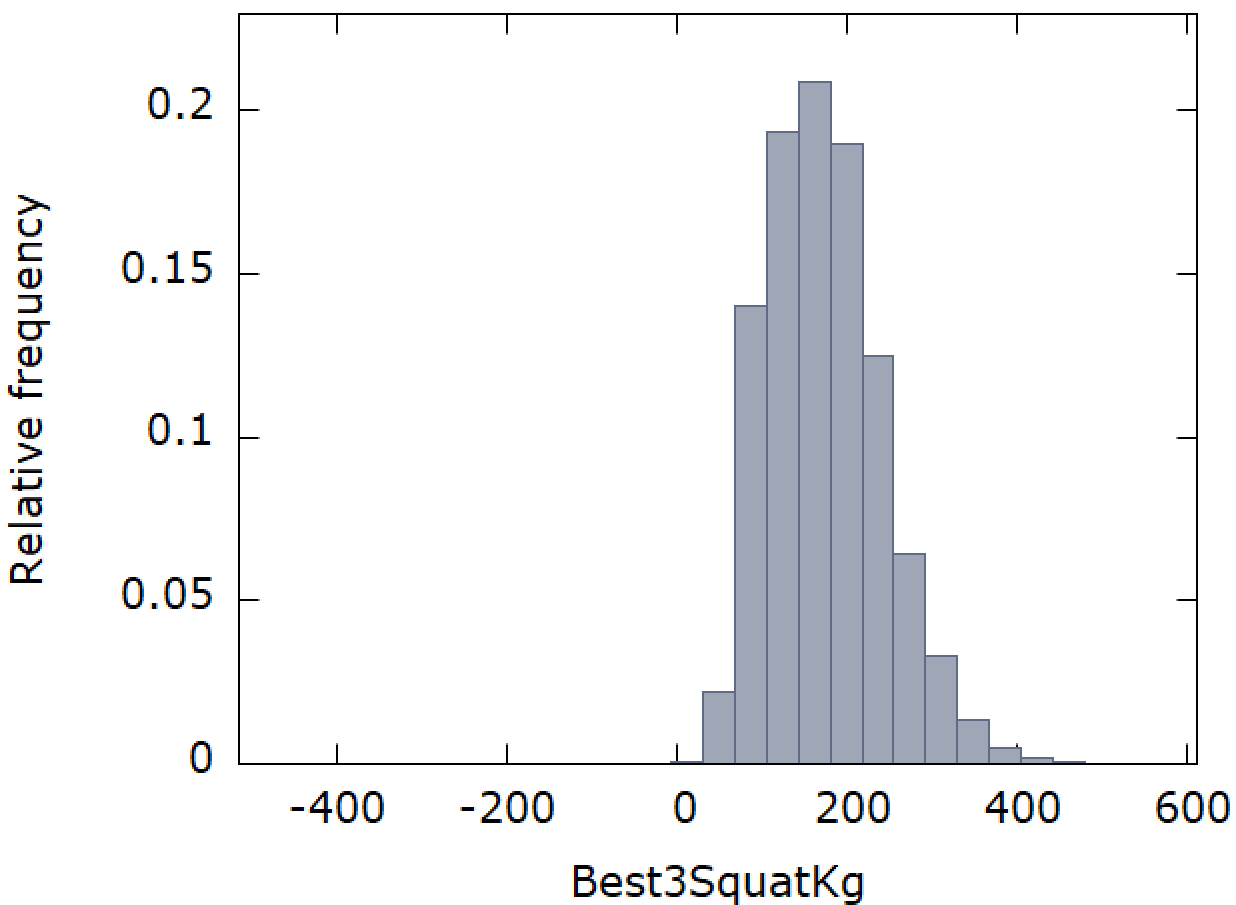
Description automatically generated with low confidence

Chart, histogram

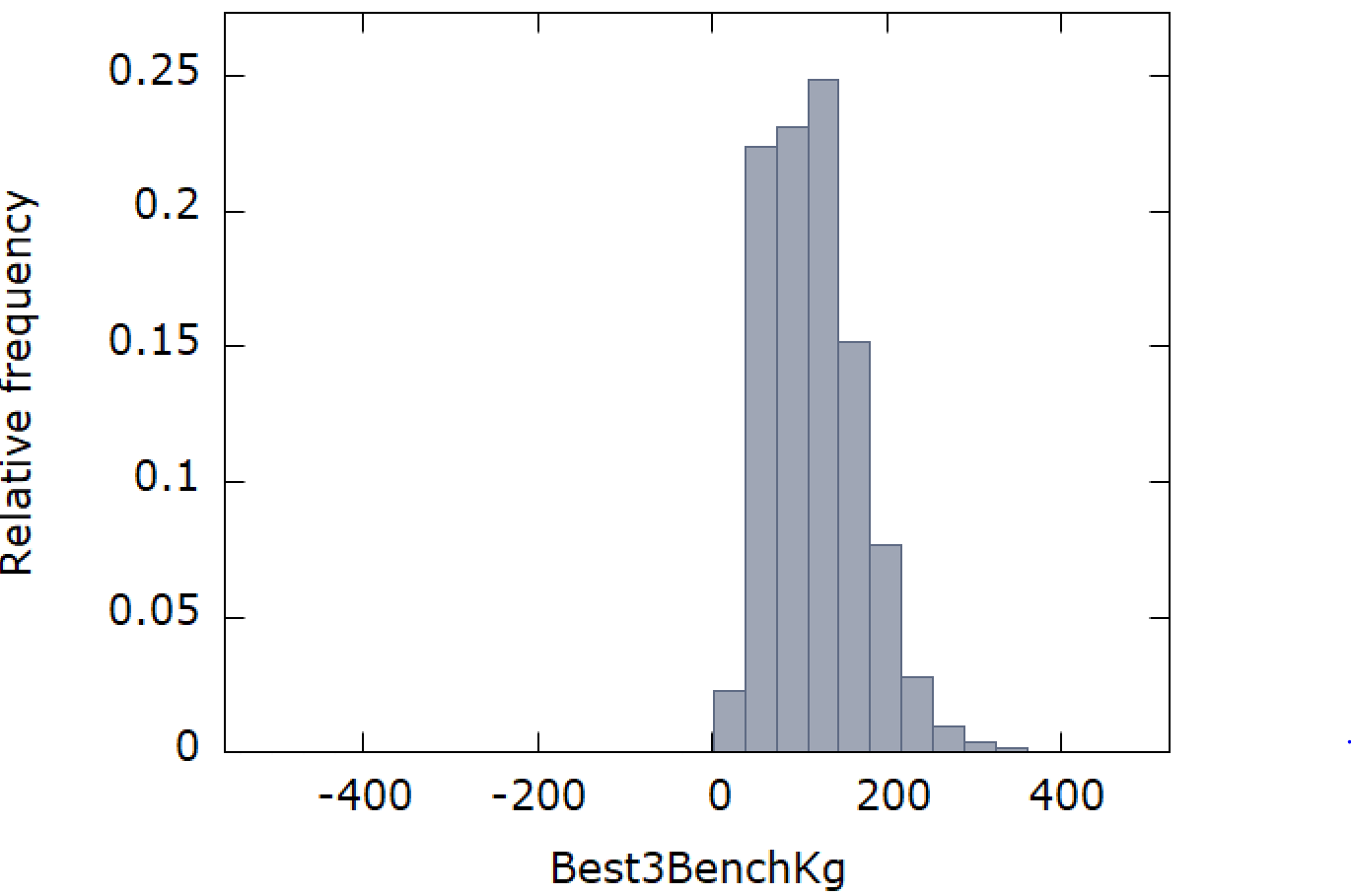
Description automatically generated

Chart, histogram

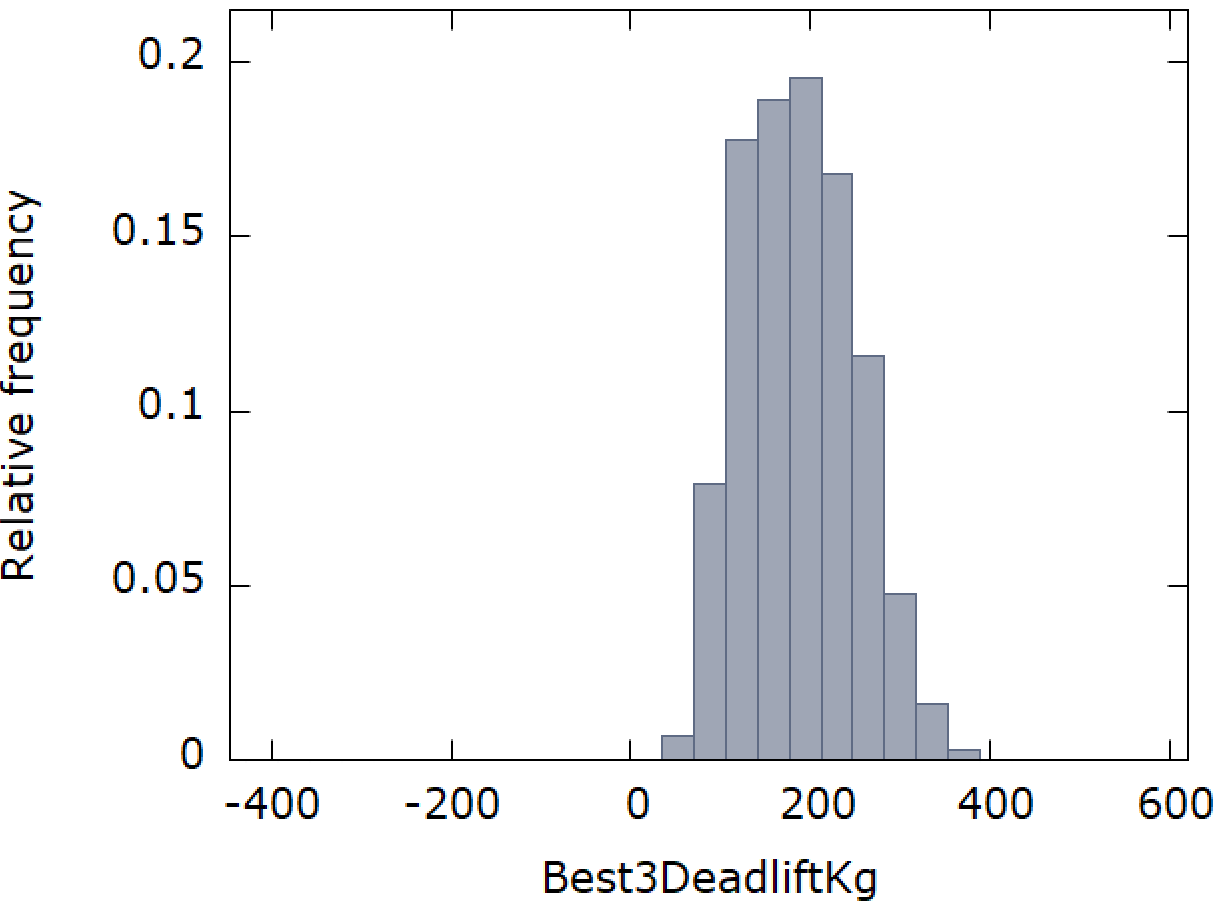
Description automatically generated

Chart, histogram

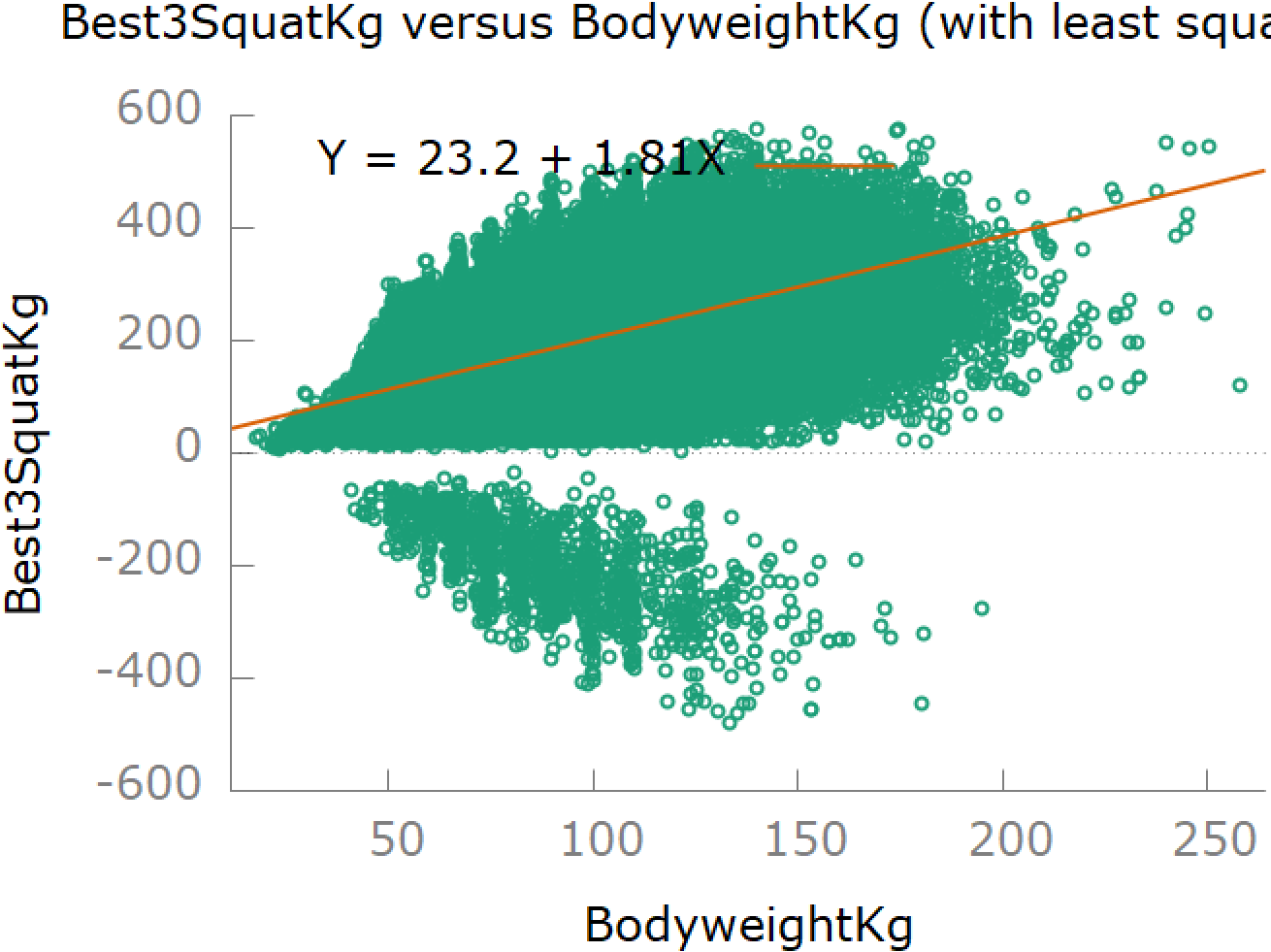
Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated





I’m going to have to either reduce the size of my dataset (clean out all missing observations and maybe take data from a smaller time period) or take random samples from my dataset to get useable correlation graphs.

I’m also considering removing failed lifts (the negative numbers). I chose “best of three” because I figured everyone would have at least one successful lift, but evidently that is not the case.