COMS4060A/7056A:

Univeristy of the Witwatersrand 2023

August 31, 2023

Due: September 21, 2023

1 Introduction

This assignment is based on the first few lectures of the course. You will be required to perform some basic data cleaning and exploration techniques on a prescribed dataset. The aim is to explore the dataset and make observations.

Submissions should include both text and code. You may use any language that you like.

2 The Data Set

Openpowerlifting.com is a giant dataset of powerlifting competition results. It is maintained by a global collection of volunteers (with an open github repo). These volunteers have worked diligently to grow and clean the data-set, however as the data-set is huge it unfortunately has some errors.

The datset can be downloaded here: https://drive.google.com/drive/folders/1hGp6FZKQrtbjOZFeg8N3u08sQ0
(For some reason moodle isn't letting me upload)
Further complicating matters the sport is notoriously fractured. Different federations use different weight classes, allow different kinds of equipment* and have different drug testing policies to name a few.

In the sport each competitor is given three attempts at each of three lifts (the squat, benchpress and deadlift). The heaviest successful attempt at each of the three lifts is added together to form a total. However some competitions involve only the benchpress (and very occassionally other subsets are used). In some of the data each attempt (successful or not is recorded) in some of the data only the best attempt is recorded. Generally an unsuccessful attempt is recorded as a negative number (so an unsuccessful attempt at 100kg would be recorded as -100).

Additionally some international federations have branches in different countries, which are technically separate.

An additional complication is that some lifters share a name (there are at least 15 John Smiths currently). In this case the name has a #n added to it for example John Smith#15.

*The most common kinds of equipment are raw, raw with wraps, single-ply, double-ply and unlimited.

3 The assignment

As the data-set is very large you won't be required to use the whole of it. Instead you should pick between one and three federations and work with these. As I'd prefer everyone to use different data-sets the South African federations (sapf, rhinopc and wpc-sa) should not be picked.

You should:

- Combine the data in your dataset (subset of the data).
- Convert all weights to a common unit (kg or lbs). This includes bodyweights and weight classes.
- For each performance (each competition of each lifter) compute columns giving the best performance on each lift.
- Powerlifting has many formulas to compare lifters of different bodyweights. There are endless arguments about which one is "best" or appropriate to which division. While you can read about the assortment of these formulas here for this assignment you need only consider the two below.
  - The wilks coefficent: Description here
  - IPF points: Description here (Use equipped for single-ply double ply or unlimied and classic for raw or raw with wraps for this assignment)
- Engineer two new features one for each of these two scores in your dataset. Show a few of the datapoints in your report.
- Compute correlations between these two features
- Compute mean, modes, variance, inter-quartile ranges and so forth for both folrmulas.
- Filter to only using a single performance for each lifter (best or most recent) and repeat the above calculations. Comment on how these two analysises differ?
- Plot lifter body-weights. What do you notice? Why is this likely the case?

- Make a correlogram between bodyweight, squat, benchpress and deadlift. Do this on your entire dataset and filter based on gender category and equipment used. Comment on the results
- The dataset is not perfectly tidy. Some lifters in the dataset are entered under multiple different names. Your assignment is to find as many of these as possible. This is very much not an exact science but some good proxies are:
  - Similar names or one name beign a shortening of another
  - Lifting in the same federation/federations or in similar geographical areas
  - Lifting relatively simialar amounts
  - Consistent ages

  Every example you find beyond the first is a bonus mark (but you cannot get over 100 percent total for this assignment).