Analysis of the Open Powerlifting Dataset

2022-06-27

Niklas L.

- TL;DR - The Bottom Line Up Front
- Introduction
- Data Exploration
  - The First Look
  - Demographics - Sex, Age, Bodyweight, Equipment and Locations
  - The Athlete’s Performance
  - Influence of Bodyweight
  - Influence of Age
  - Bringing Both together - Age and Bodyweight
- Statistical Testing
  - Tested vs. Untested Competition
  - Athletes Competing at Home vs. Athletes Competing Abroad
  - Differences Between Classes of Equipment
- Clustering - Are There Different Types of Lifters?
- Prediction Tool - Random Forest (Spoiler: Pretty Mediocre)
- Summary

TL;DR - The Bottom Line Up Front

- The sport of powerlifting grew substantially, especially in the last decade
- A sizeable portion of this growth can be attributed to a steep increase of female athletes, who made up 25% of all lifers in 2019
- COVID-19 resulted in a massive reduction of competitions, a hit the sport still recovers from
- Competing women’s mean age and bodyweight increased slightly during the sports growth period, settling at 32 years and 72KG respectively. Competing men’s mean age oscillated between 30 and 32 years, while the mean bodyweight remained remarkably constant at 90KG
- Weight cutting or at least deliberate attention to one’s bodyweight seems to be common. The various weight classes upper limits can be made out in a density plot of the athlete’s bodyweight
- Equipped lifting lost it’s grip on the sport, reducing it’s ‘market share’ in a massive manner since 2007, while raw powerlifting became by far the most popular category of

competition

- Most meets took and still take place in the US, followed by (especially eastern) Europe
- Competition became steeper as the sport grew, especially in the women's divisions
- Unsurprisingly, bodyweight is massively influential for all lifts, though to varying degrees
- Age is another influential factor, with peak performances being delivered between 24 and 39 years, even though incredibly impressive totals are achieved in all age groups
- There are significant differences between tested and untested competitions (38KG for men, 9KG for women with the untested athletes being stronger), as well as between athletes competing at home and those competing abroad (38KG for men and 21KG for women with those competing at home being stronger). However, this might very well be due to differences in variables which I did not account for
- Equipment does have an impact, though the results are not as clear as one would expect. Data is also lacking, especially for some of the women's divisions
- The athletes can be clustered depending on how the individual lifts contribute to the athlete's total. This offers some interesting hypotheses. Differences in age, bodyweight and strength between these clusters exist
- Predicting an athlete's total is challenging when using a simple random forest model

Introduction

This analysis was designed to be the final assignment for the Google Data Analytics Professional Certificate offered by Coursera. It's purpose was to gain interesting or valuable insights while practicing various analytic skills in R, including data cleaning and wrangling, various plots and visualizations as well as different statistical testing and modeling techniques. While I do not anticipate many readers outside of my peer graders, I would still like to point out to the odd reader who stumbles across this article that I am thankful for any kind of feedback. In case you notice any mistakes, parameter misinterpretations or have other suggestions for improvement: Feel free to let me know!

This analysis will deal with the topic of powerlifting, a strength sport consisting of three lifts: The back squat, bench press and deadlift. When competing at a standard powerlifting meet, each lifter will be given three attempts at maximal weight on each of these three lifts for a total of nine possible attempts per meet. The single best successful attempts in squat, bench press and deadlift will then be added up to form the lifters *total* , which in turn determines the athlete's rank.

I will work with data obtained from the open powerlifting project, which tries to create a comprehensive and accessible record of lifts performed in powerlifting meets all around the world. The main objective here is twofold:

• First, to shed a little light on the sport as a whole. Has powerlifting changed in the last few decades? If so, what exactly are the changes we can observe and what are their likely causes?

• Second, to find out more about the backgrounds of the athletes performance. While there is limited information about how exactly an athletes performance came to be, I still want to explore patterns and discuss some possible influences and the extent of their impact.

To get this done, I will first explore and visualize the data in respect to a variety of key topics before moving on to conducting a few statistical tests for interesting questions that arise during the exploration. During the course of this endeavor, I will focus mainly on discussing the results and less on the code I wrote to obtain them. If you are interested in the code, feel free to take a look at the original file provided in the repository.

So, let's get started!

Data Exploration

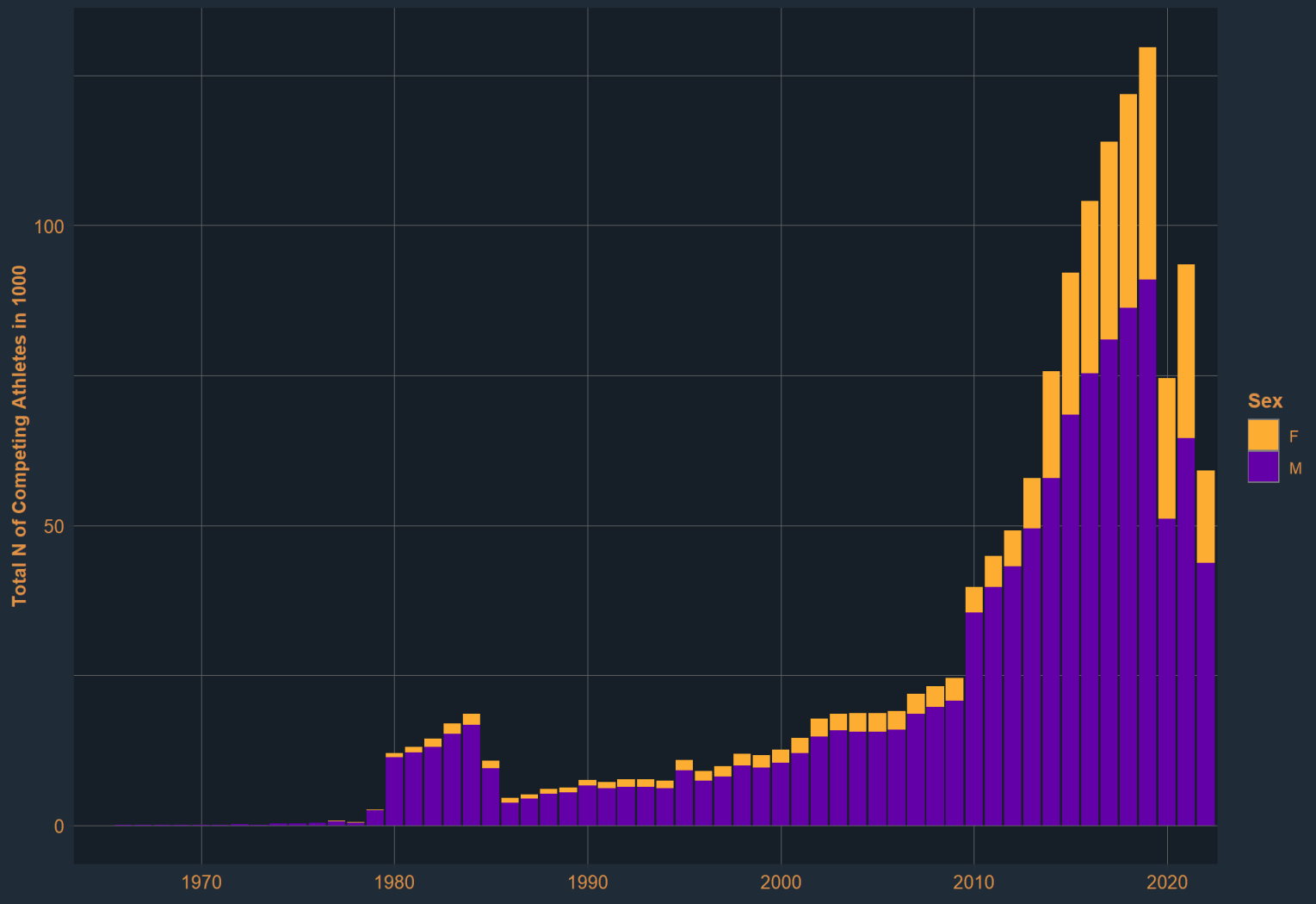
The First Look

The dataset includes the competing athlete's name, sex, age and bodyweight as well as information about the meet and each attempt the lifter took. We can also check for the equipment-class the athletes decided to compete in as well as whether or not the competition was drug-tested or not. Conveniently most of the columns names are pretty self-explanatory, which makes additional renaming unnecessary. Also there are some columns of which I already know that they will be of very little use during the course of this analysis, like the federation under or specific town in which the meet took place. After removing these columns, we are left with 32 columns and roughly 2.7 million recorded performances. Quite a lot of information!

```
##  [1] "Name"            "Sex"             "Event"           "Equipment"
##  [5] "Age"             "AgeClass"        "BodyweightKg"    "WeightClassKg"
##  [9] "Squat1Kg"        "Squat2Kg"        "Squat3Kg"        "Squat4Kg"
## [13] "Best3SquatKg"    "Bench1Kg"        "Bench2Kg"        "Bench3Kg"
## [17] "Bench4Kg"        "Best3BenchKg"    "Deadlift1Kg"     "Deadlift2Kg"
## [21] "Deadlift3Kg"     "Deadlift4Kg"     "Best3DeadliftKg" "TotalKg"
## [25] "Place"           "Wilks"           "Tested"          "Country"
## [29] "Federation"      "ParentFederation" "Date"           "MeetCountry"
## [33] "MeetName"        "Year"
```

Demographics - Sex, Age, Bodyweight, Equipment and Locations

Next up, I believe it makes sense to learn a little more about the competitors whose performances we are looking at here. First question: Has the number of competitors changed over time, especially in respect to the proportions of sex?



The first and most obvious takeaway here is that the number of athletes competing in powerlifting meets increased quite substantially over time, with the most notable growth happening between 2010 and 2019, the latter being the single year in which we can see the highest number of competitors yet.

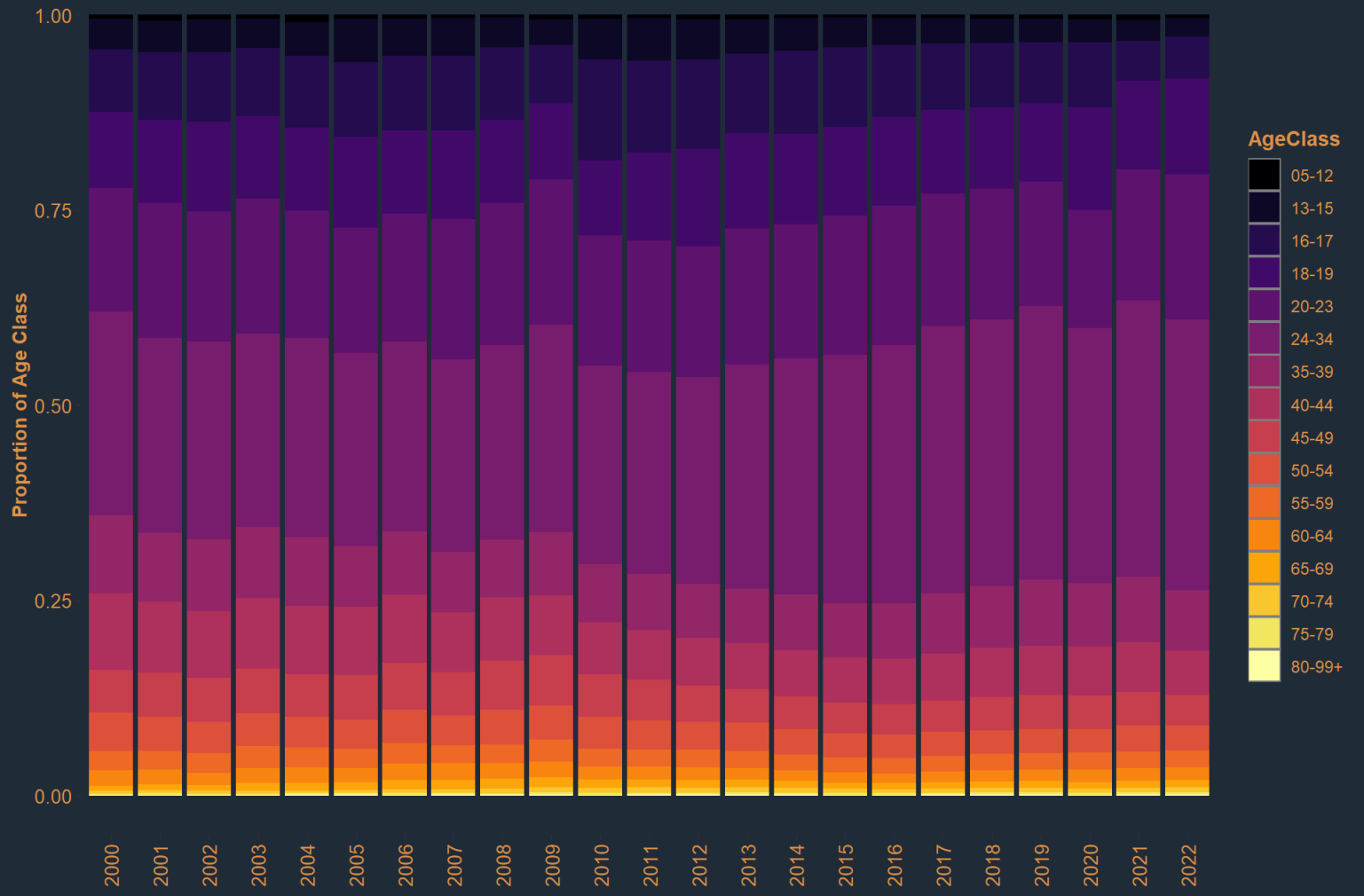
A second key point of information is that this growth seems to be at least in part due to a notable increase in female athletes, whose group registered an especially impressive growth in between 2014 and 2019. Their counterpart in the population of male athletes is not far behind, however, also registering strong and steady growth during the last decade.

An interesting point of note is the sharp decline in total competitors in 2020, which is likely due to the impact of the Coronavirus. As we can clearly see, powerlifting did not manage to elude the consequences of the pandemic with the total number of competitors dropping from above 125,000 in 2019 to just below 75,000 in 2020. There seems to be a trend towards recovery in 2021, however, which can hopefully be taken as an indication that the sport can manage to return to its pre-pandemic trend of growth before long. Making strong statements about the state of powerlifting in 2022 is hard, since we only have access to roughly half the years data as of now.

The final bit of information that can be pointed towards is the lack of data during the earlier stages of the sport. As fascinating as the history of the sport can be, the data collected before the turn of the millennium is of lower quantity (and often, sadly, quality), which prompted me to exclude it during the rest of this analysis. From now on, I will limit this analysis to data

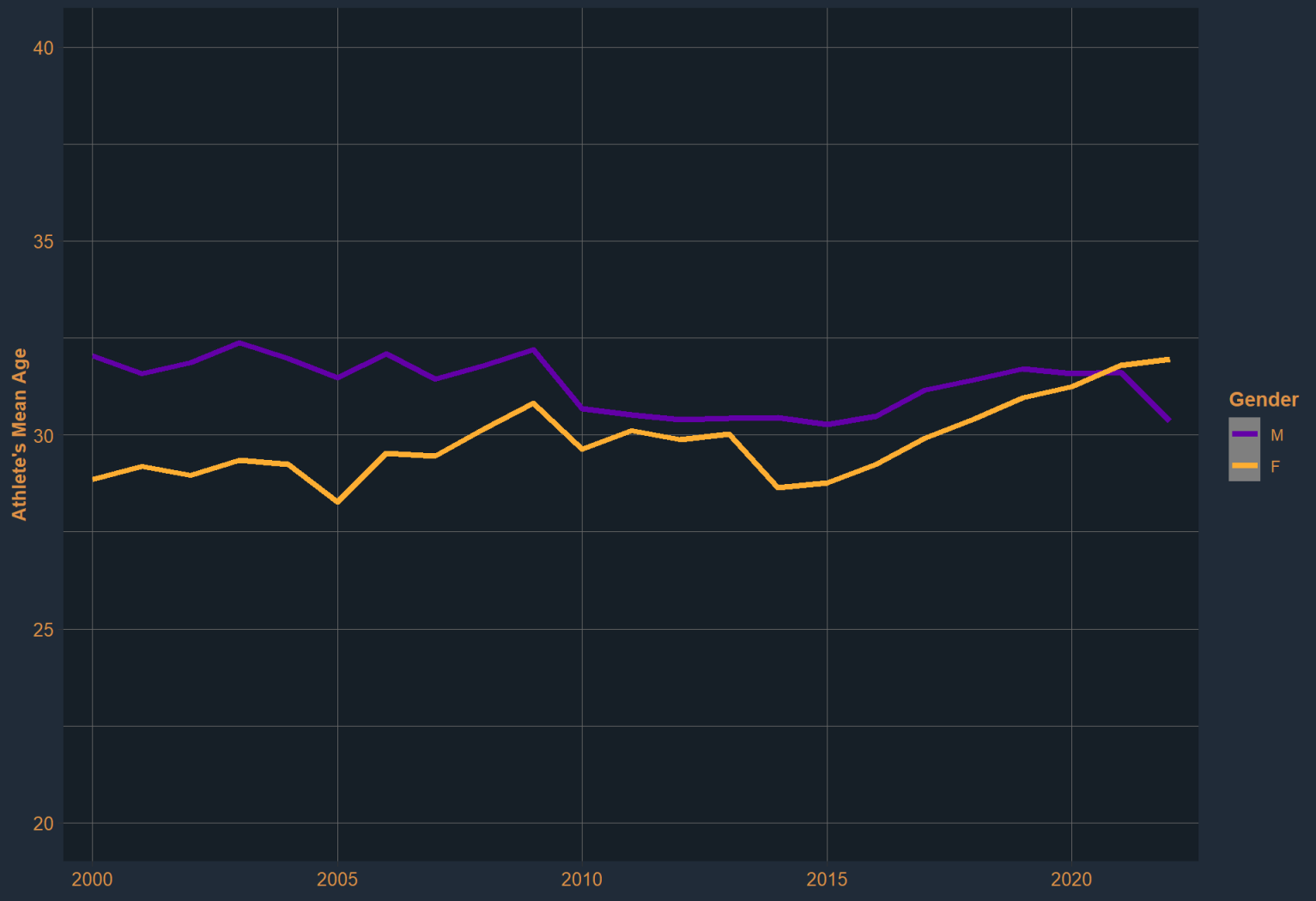
collected from 2000 onward. I will also limit my analysis to lifters of male and female sex, since this makes visualizations a lot less cluttered and easier to understand. I do not mean to exclude anyone, but since the number of lifters in the dataset that do not fall into either the male or female category is miniscule, no reliable information can be gained here.

That being said, let us continue to the second question of interest during the exploration of demographics: The competitors distribution over the various recorded age classes.



There is not a whole lot of change to see here. In general, the distribution of competitors across age classes seems to remain relatively constant, which is worth taking note of when considering the rapid growth and massive influx of new lifters the sport went through during the last decade. The most notable change we can observe is the slight growth in medium age groups, especially the group of 24-34 year olds, who managed to cement their lead and still constitute the biggest group of competitors in 2022.

As an additional measure to confirm this trend, or lack thereof rather, Let us plot the competitor’s mean age over the last two decades:



What we already saw in the plot of age classes above can be, to a degree, validated here: While there are no dramatic changes in the mean age during the last two decades, there are some nuances we can notice. The men are peacefully oscillating between 30 and 32.5, dropping somewhat in years of pandemic, while the women managed to close the previously existing gap with a noticeable increase in mean age between 2014 and 2021, surpassing that of the men in 2022.

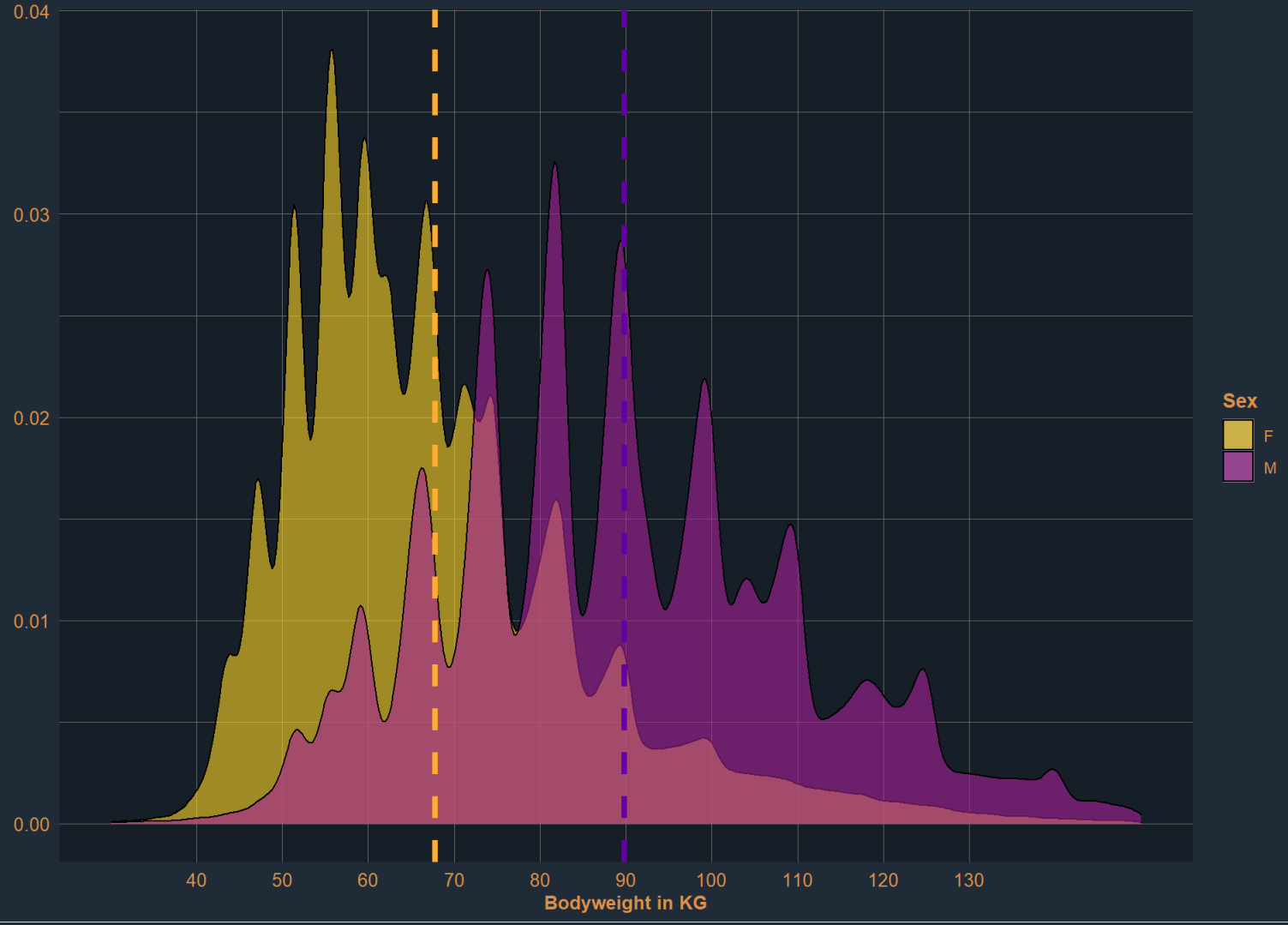
Now that we shed a little more light on sex and age, the next factor I would like to take a look at is the athlete’s body weight. This is often considered an especially meaningful metric, since being heavier allows an athlete to carry more muscle on their frame, which can have a huge impact on performance. I know. I was shocked to learn about this as well.

So, let us take a look at whether or not the mean bodyweight of athletes changed over the last years:

Slight increase over time for the male population, more noticeable increase in mean weight for the female population, from slightly above sixty to above seventy kg after 2020, an increase of roughly ten KG!

What we already saw in the visualization of the athletes age holds true for their bodyweight as well: The men’s data remained remarkably constant with curve tightly hugging the 90KG line. The mean bodyweight of the women on the other hand increased by roughly 10KG, from a little abve 60 to slightly above 70KG. This is a substantial increase, enough reason to explore the topic of bodyweight a little further.

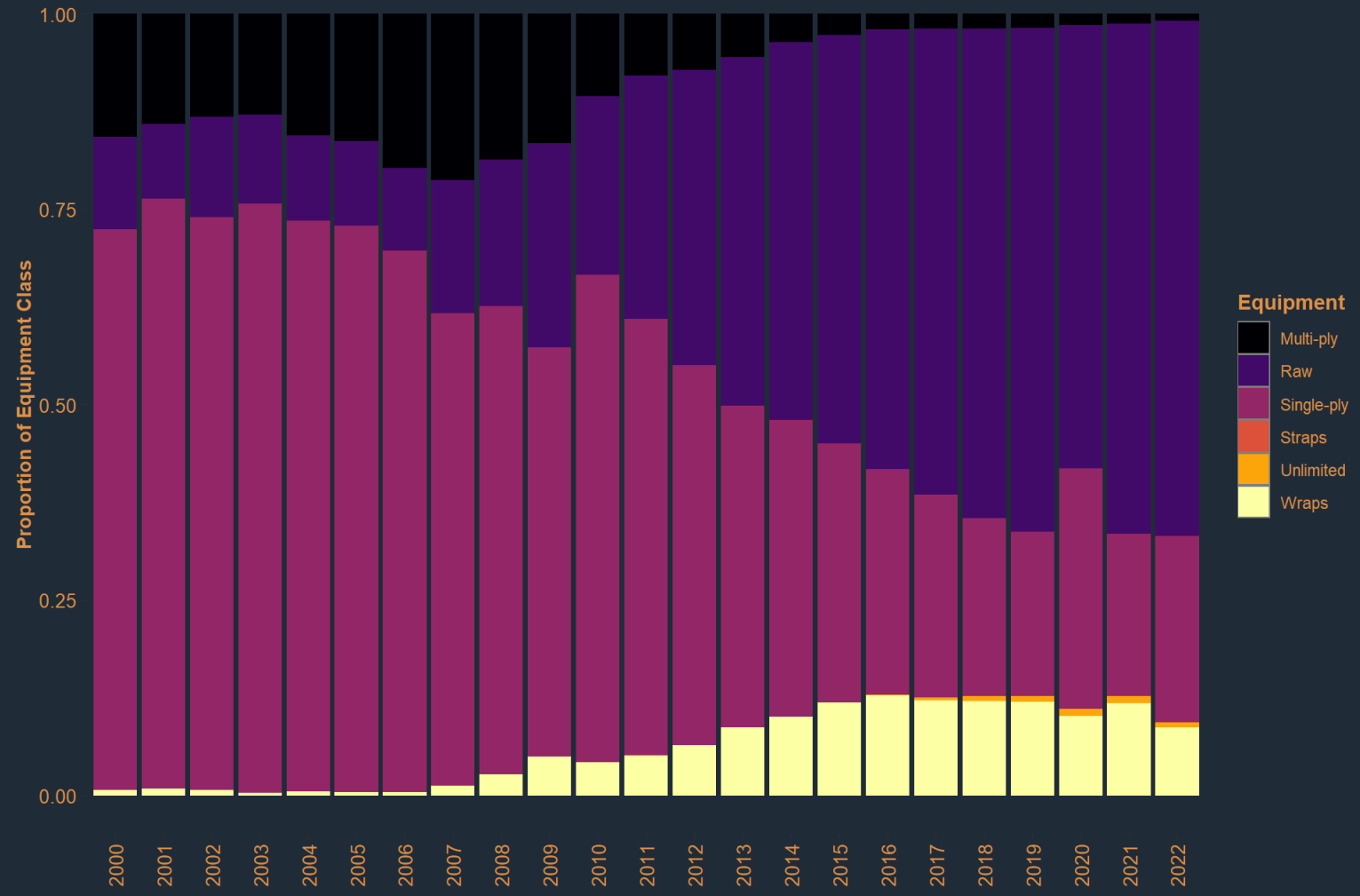
So, let us take a look at how exactly these mean values come into existence by plotting a distribution of the athlete’s bodyweight. How is the athlete’s bodyweight distributed around these mean values we can observe here?



This is an interesting result. While the weight distributions looks somewhat normally distributed and the bulk of competitors fall around the mean values we observed, we can also notice sharp peaks in the distributions, making them look pretty unnatural. An obvious explanation would be that what we are observing here is the alignment of bodyweight to the various weight classes in the sports different organizations. In the female population we can, for example, observe an accumulated amount of lifters around both 52 and 57 KG, two of the more popular weight classes in the IPF, one of the more impactful powerlifting federations (more on the different weight classes can be read here).

While this gravitation towards certain thresholds seems to be of little surprise, it can also be taken as an indication that weight cutting or at least deliberate control of one's bodyweight in preparation for a competition is a widespread practice in competing athletes.

Now that we have explored most of the more obviously influential factors like gender, age and body weight, there is another notable aspect in which competing athletes should be distinguished: The equipment they use in competition. Powerlifting competitions can allow various forms of supportive equipment, from knee-wraps to supportive suits, all of which can aid the lifts in one way or another. This can make quite the difference in what weight an athlete is able to successfully lift (more information on that can be found here). Taking a look at which of these categories are the most popular should be worth the time:



Now this is quite an interesting result!

Both Single and Multi-ply, the two main categories in which supportive equipment is used, lost in popularity as the sport grew, while Raw, the category in which only kneesleeves and wristwraps are allowed became by far the most popular category in the last decade. The Wraps-category, in which knee-wraps are allowed as well, seems to have carved out a moderate niche as well despite being virtually nonexistant around 2005.

The exact reasons for these changes are hard to confirm, but I believe that an increased accessibility of the sport might play an important role. As we can see, lifting in supportive equipment of some kind was the most popular form of powerlifting after the turn of the millenium. This equipped lifting is far less accessible to new athletes, since getting used to, or even just buying a lifting suit or bench shirt might have been a steep barrier for many new athletes. Competing raw, in comparison, is much closer to what usually happens in the gym. If I were to show up at a given meet in sweatpants and a shirt, I would be placed in the raw category. Again, this is only speculation and the direction of causality is hard to assess from this data alone, but even if my explanation is not correct, this trend is another interesting aspect of the sports development.

Also, while we are discussing the importance of equipment anyway, I would like to point out how much of a factor this can be in practice by taking a look at the maximum bench press performed in a few categories:

```
##   max_bench_raw_kg
## 1              355
```

As we can see, the highest amount ever successfully lifted in the raw bench press clocks in at 355KG, which is an absolutely incredible number already. Still, the record in the ‘Single-Ply’ category manages to dwarf this lift, currently residing at an incredible 508KG.

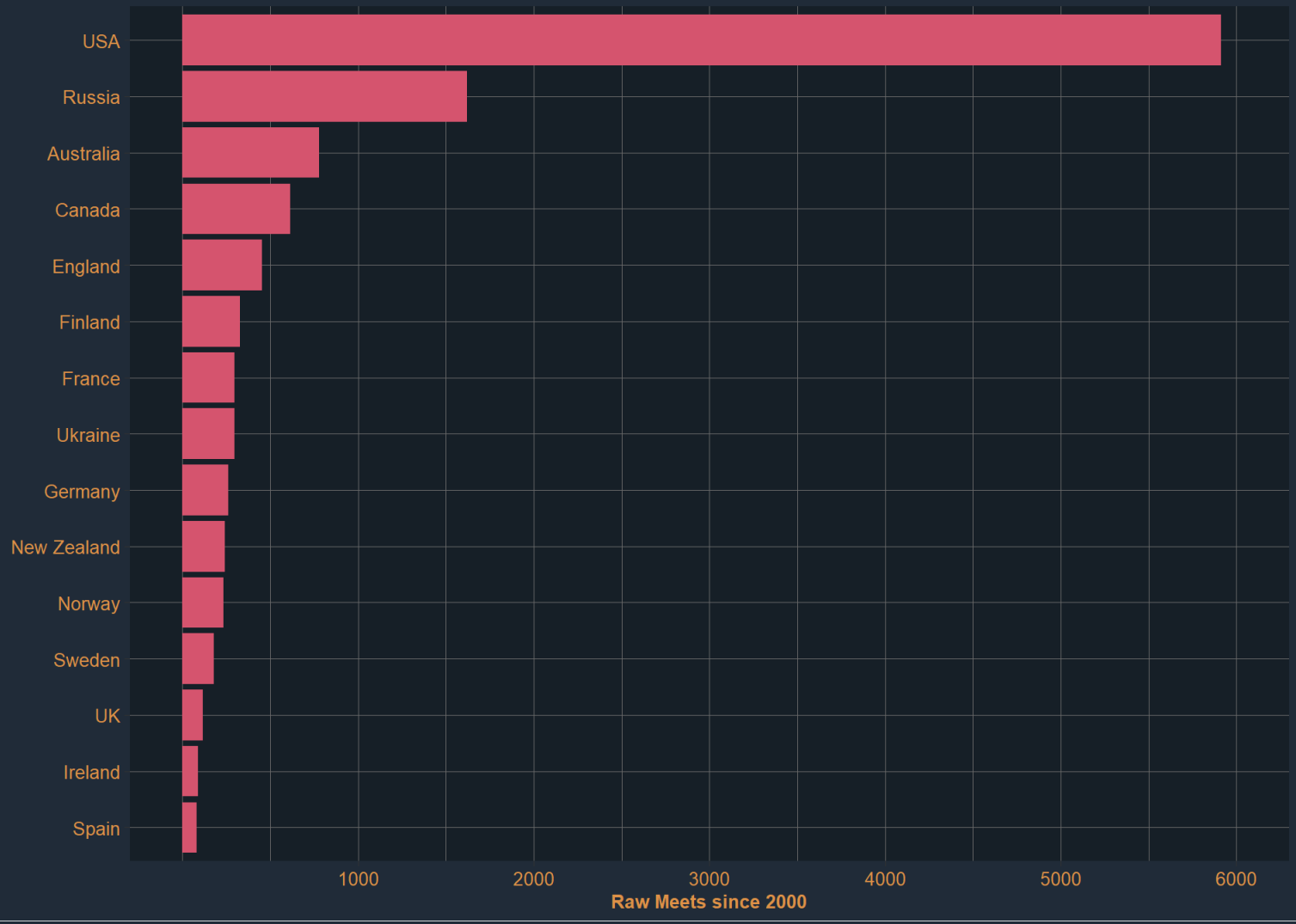
```
##   max_bench_single_ply_kg
## 1                  508.02
```

But we can take this even further. When taking a look at the, admittedly, somewhat niche category of ‘Unlimited’, the weights become outright absurd, with the biggest bench sitting just short of 600KG.

```
##   max_bench_unlimited_kg
## 1                598.74
```

Clearly, the usage of equipment plays an important role in what amount of weight an athletes is able to successfully lift and accounting for these differences in the parts of this analysis yet to come might be more confusing than enlightening. I will return to the impact of equipment later on, but will in the meantime focus the analysis on the category which seems to be the (main) cause of the sports increase in popularity: Raw Powerlifting. I hope this makes the parts of this analysis yet to come more comparable to what the average reader could spot in the gym, while preventing confusion about what class we are talking about at any given moment. I know that reducing the amount of data that is used in the analysis is not ideal, but fortunately this dataset is large enough to offer the luxury of focussing on a specific subgroup without risking the value of our results.

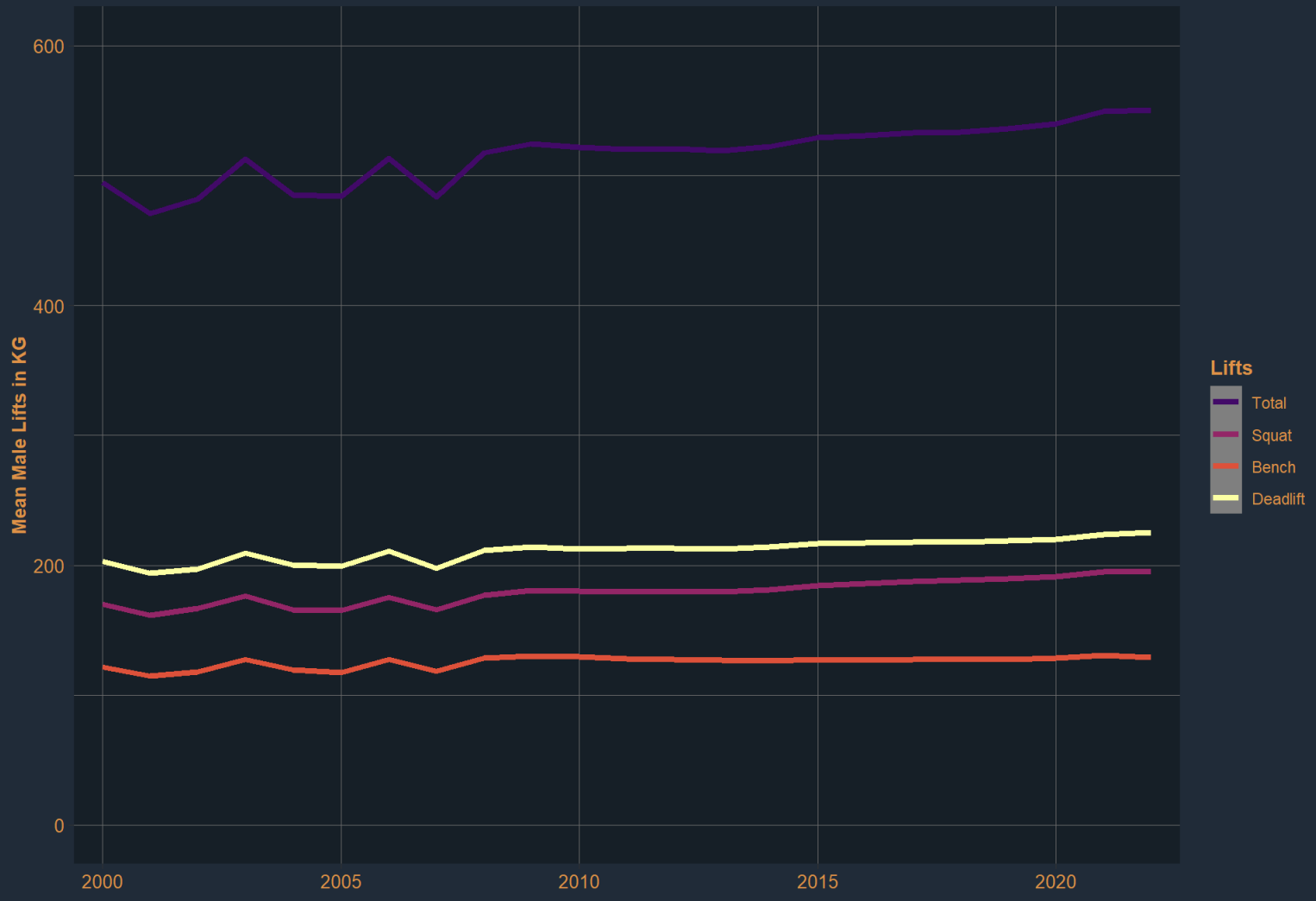
That being said, I want to move on to another interesting aspect of the sport: Its popularity in different parts of the world. To get an idea of the sport’s popularity in different parts of the world, let us plot where most competitions took place:



Well, a pretty clear favorite to be spotted here. Most of the Raw meets during the last two decades took place in the US, with no other country being remotely close. However, (especially eastern) Europe and Oceania seem to be doing alright as well, while no African or Asian country manages to place in the top 15.

The Athlete’s Performance

Now that we know a little more about who competes in powerlifting, what weight these athletes compete at, what equipment they use and where the competitions take place, it is time to get to the core of the dataset: The athlete’s performances. Since the differences between male and female competitors are pretty notable, I will divide my visualizations for this topic in an effort to do both groups justice. So, let us start by taking a look at how the performance in the total as well as it’s three components changed over the last two decades, starting with the mean performances of the male athletes:



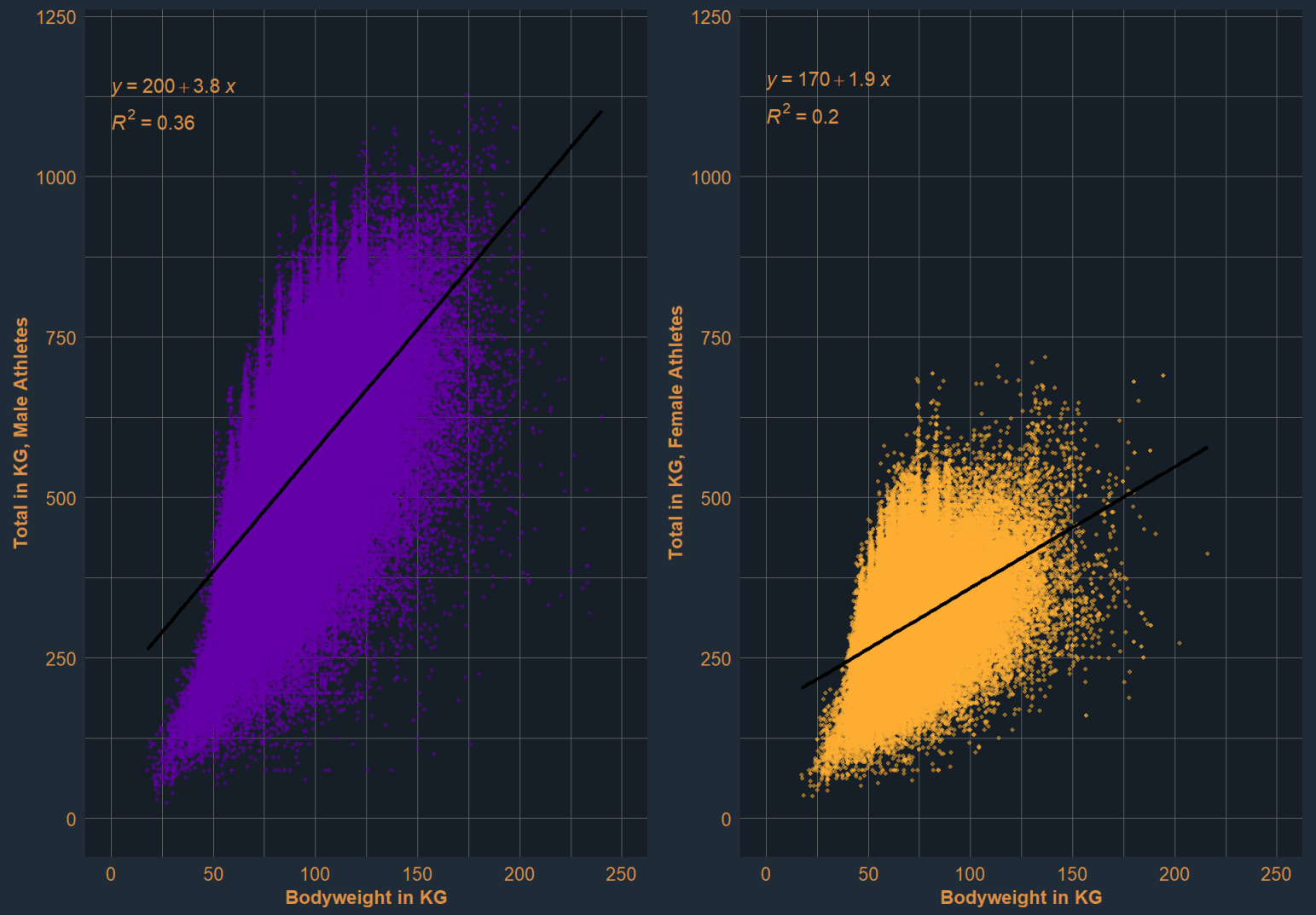
The men continue their trend of not really changing a whole lot about what they are doing. Still, we can notice a modest increase in the total over the last two decades, which now manages to reach around 550KG while sitting just below 500Kg in 2000. This increase seems to be mainly due to slight improvements in both the Squat and the Deadlift, while the bench press retained it’s mean since 2010.

If the previous insights can be any indication, we can expect more drastic changes in the women’s population. Let us take a look:

Again, quite an impressive result. The women managed to increase their average total by close to 100KG, nearly doubling the absolute increase we observed in the male population. Considering that the average total in the women's population sat slightly above 200KG in 2000, this is an increase of dramatic magnitude! A similarity to the male population are the sources of this improvement: Solid increases in both the squat and the deadlift. But while the men's bench press stagnated over the last decade, the women were able to record a modest increase.

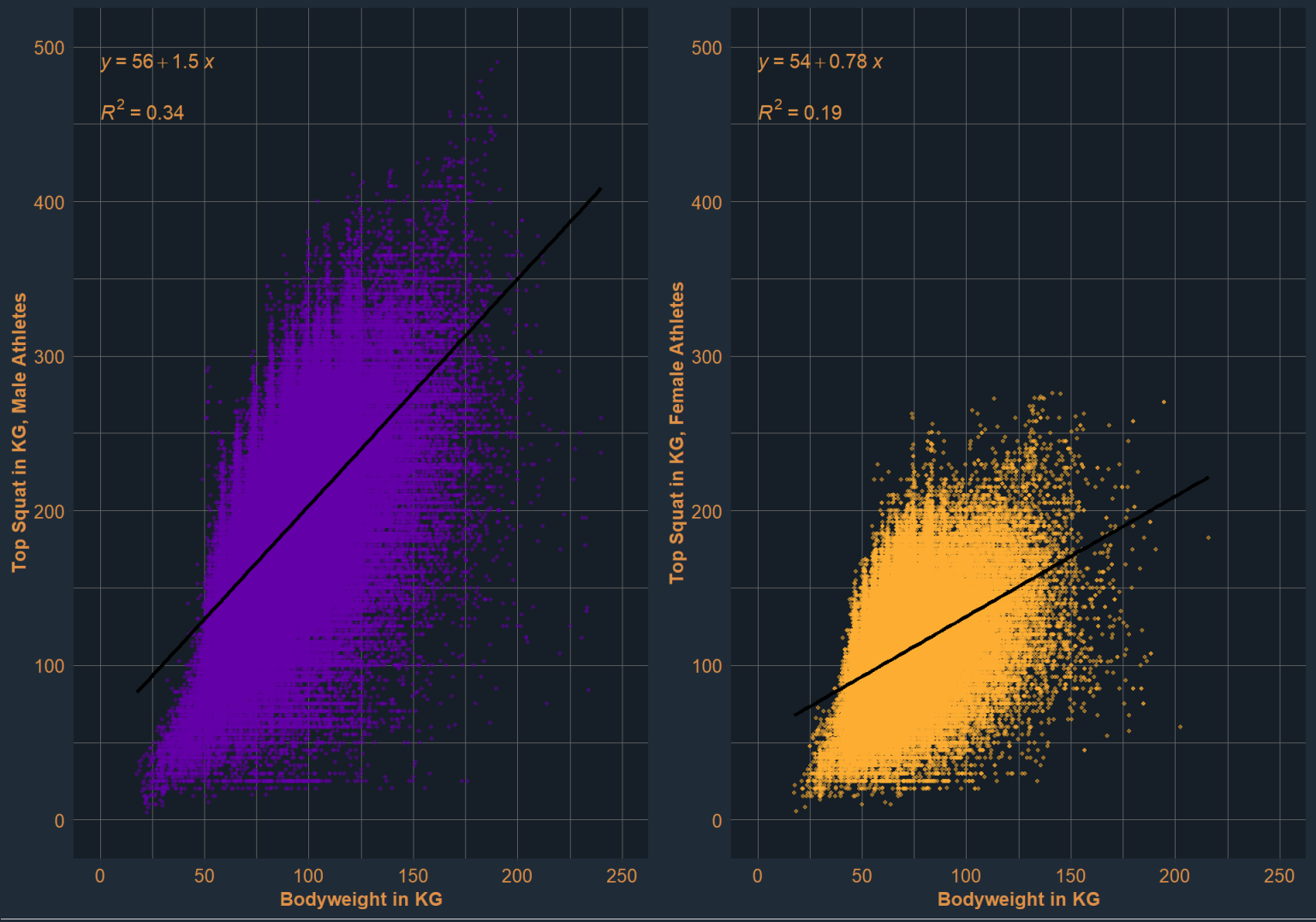
Influence of Bodyweight

Now that we know about where the athletes are at in respect to their performance, I would like to delve a little deeper into it's origins. We already explored the distribution of age and bodyweight, both of which can be assumed to be influential factors in an athlete's performance. Now I would like to know how meaningful both of these metrics are exactly, starting with the athlete's bodyweight and its importance for the total and the individual lifts. To do this, I will create a scatterplot of athlete's performances across different bodyweights and calculate a simple linear regression, which we can then use to estimate the bodyweight's impact.

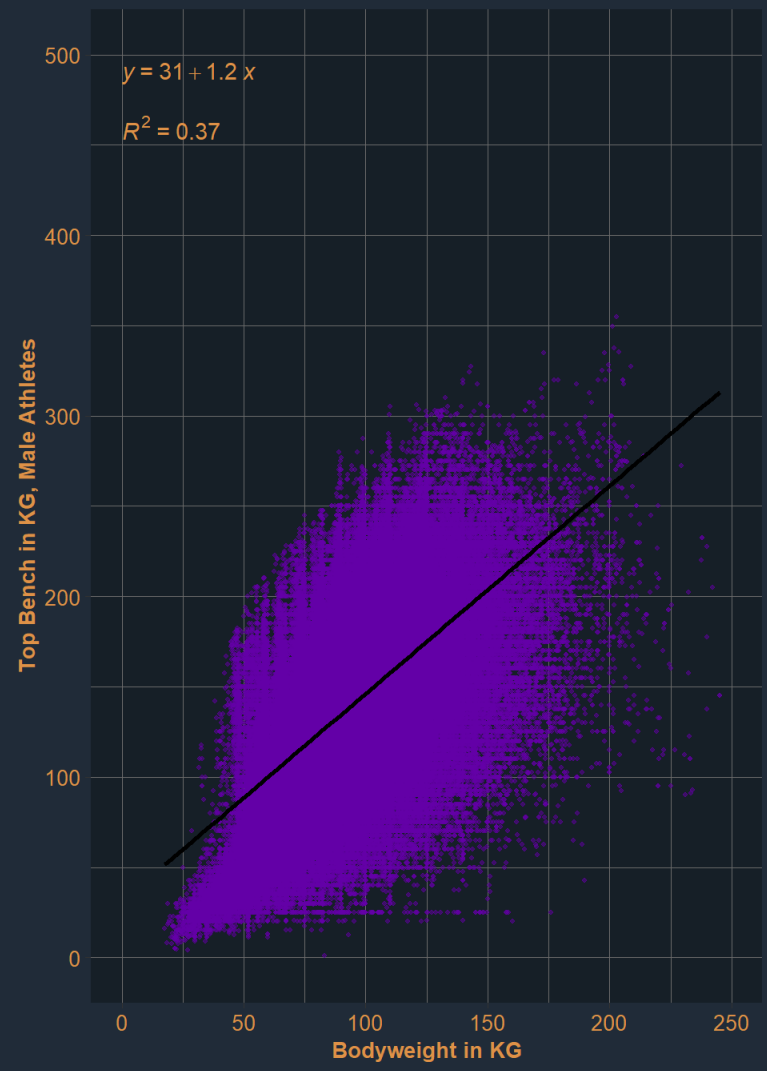


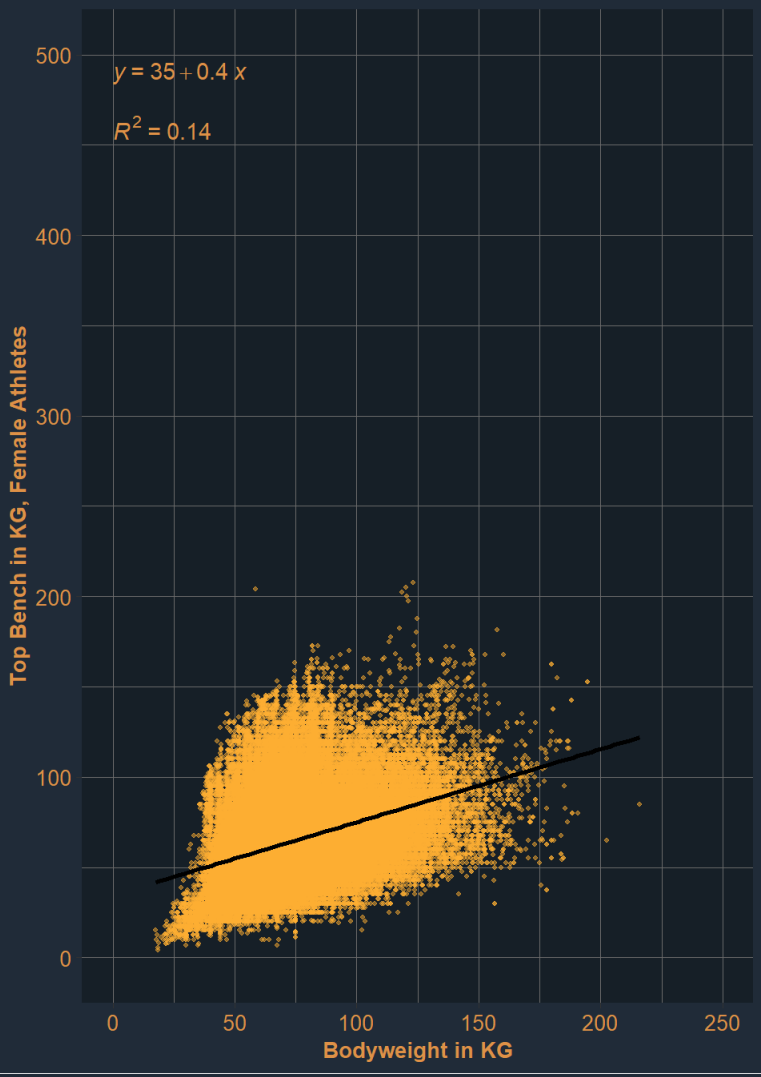
There is quite a bit of information to be taken away here. First of all, we can take a first look at the spread of totals across bodyweight for both men and women. We can notice a clear correlation between an athlete's bodyweight and their total, indicated by the black line. The steeper this line, the more influential the bodyweight for our outcome metric (the total in this case). We can use the slope of this regression line to make an estimate as to how impactful the bodyweight is: With an increase of 1KG in a male athlete's bodyweight, we can expect a gain of roughly 3.8KG in the total, while a female athlete can expect approximately half of that with an increase of 1.9KG per additional KG of bodyweight. In total, the differences in bodyweight are able to explain 36 percent of the variance in the men's but only around 20 percent of the variance in the women's total.

We can apply the same procedure to the three individual lifts as well, so let us start by taking a look at the relation between bodyweight and squat performance:



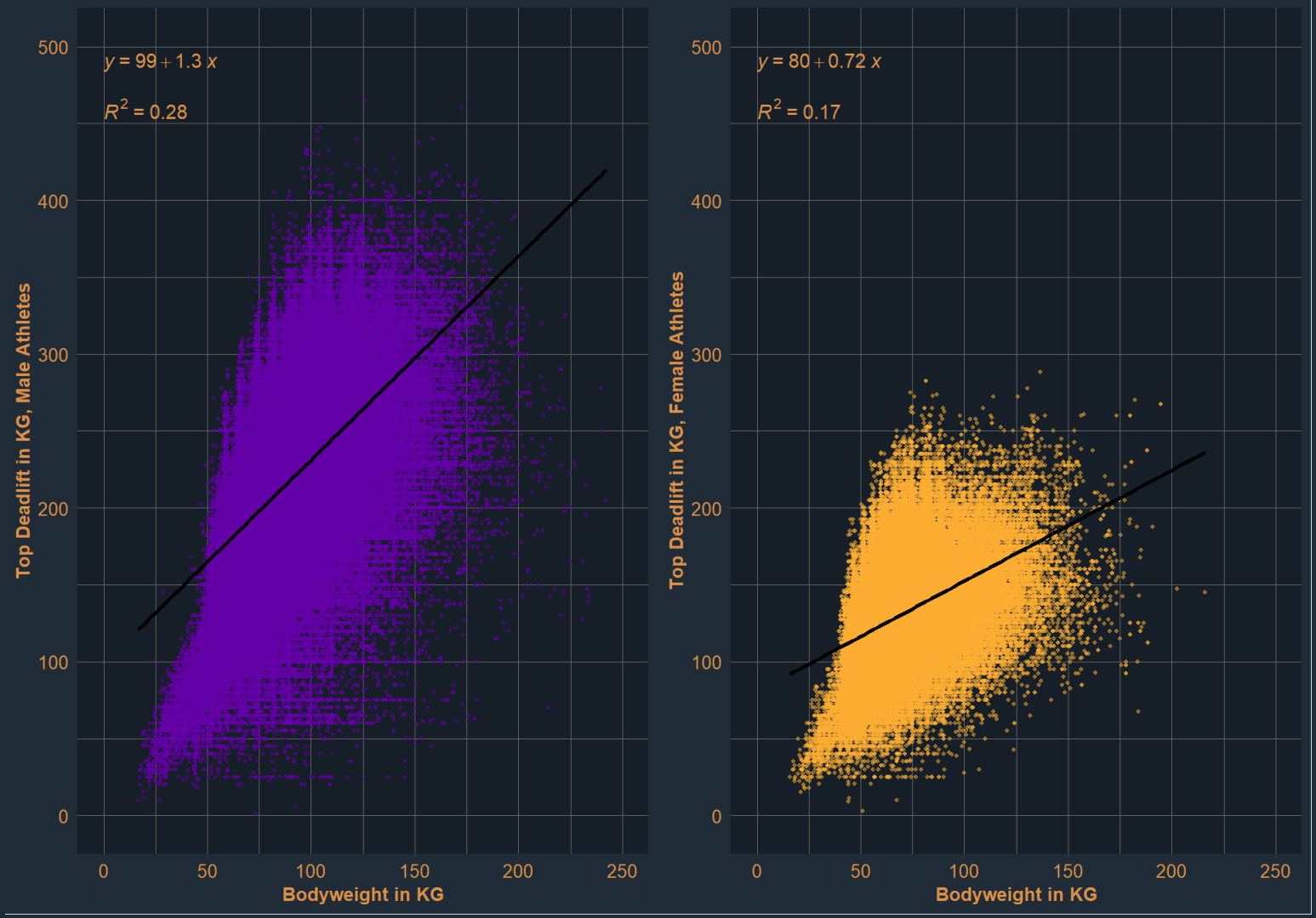
Applying the same procedure to the performance in the Squat, we can notice a similar pattern. The spread of the data looks rather similar to that of the totals in the plot above. Bodyweight accounts for similar percentages of explained variance with 34 and 19 percent of all the variance we observe in the squat performance in the male and female population, respectively. An increase of 1Kg in the bodyweight goes hand in hand with gains of 1.5Kg for the men's and 0.78KG for the women's squat. While the amount of gain per KG of bodyweight is (obviously) lower than that we observed in the total, the explained variance for both the total and the squat is remarkably similar in both the men's and women's group. Let us see if this pattern continues with the bench press:





The answer seems to be ‘It depends’. For men and women, bodyweight accounts for 37 and 14 percent of variance, respectively. While this does not indicate a significant deviation from what we saw in the total and squat for the male population, the explained variance in the female population is noticeably lower compared to the previously observed 20 and 19 percent of variance that bodyweight explained for the total and squat. The bench press seems to be less dependent on bodyweight for women! Taking a look at the slopes of our regression lines confirms this suspicion: An increase of 1KG in bodyweight only goes hand in hand with an increase of 0.4 in the bench press for the women’s population. For the men, in comparison, we can observe an increase of 1.2KG in comparison.

Now, let us finally take a look at the deadlift:

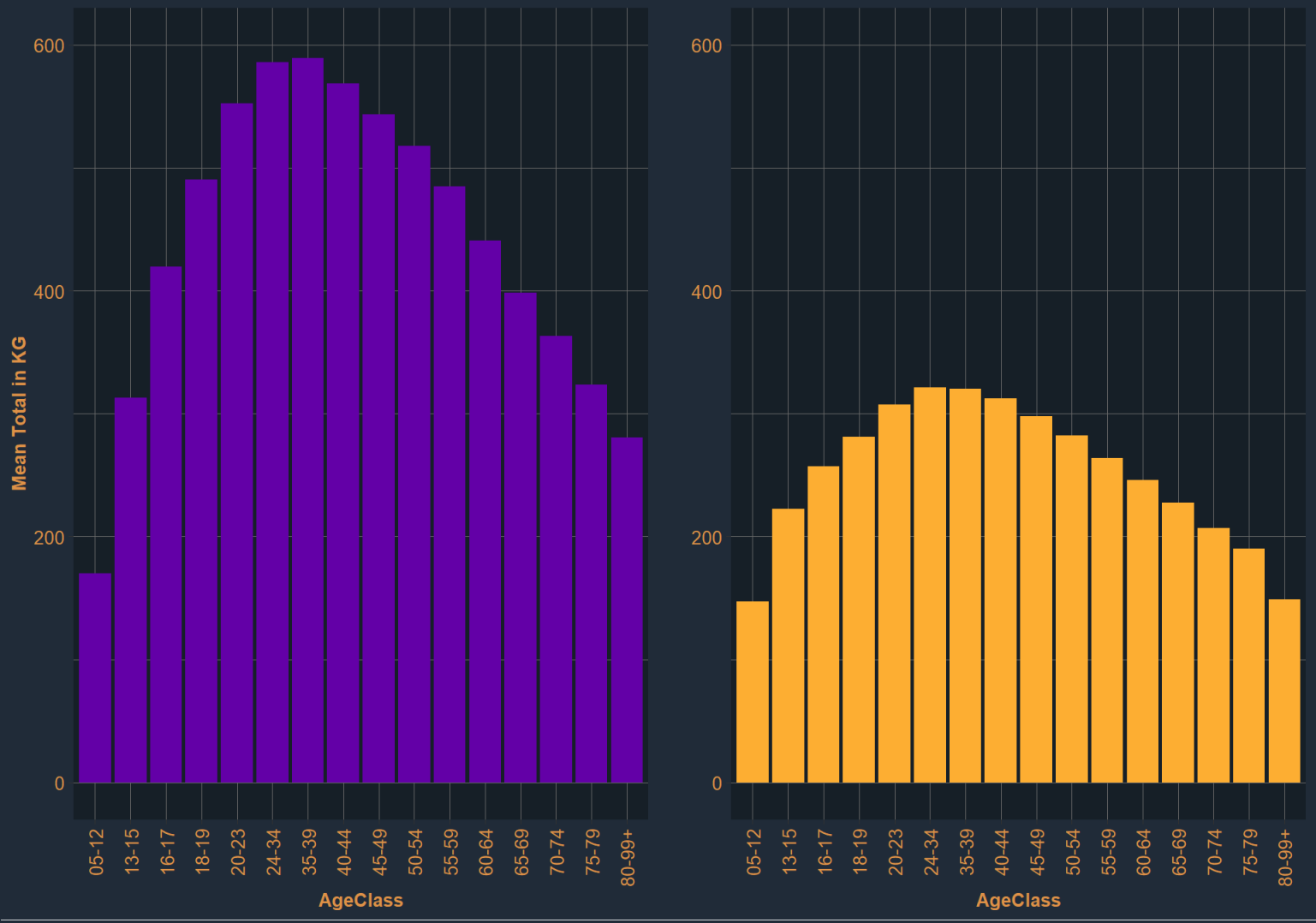


The deadlift looks interesting as well: The explained variance bodyweight can account for is relatively small for the male and more average for the female population, explaining 28 and 17 percent respectively. This is the lowest explained variance we have seen for the men so far. The explained variance in the female population sits at 17 percent, comfortably between the bench (14) and squat (19).

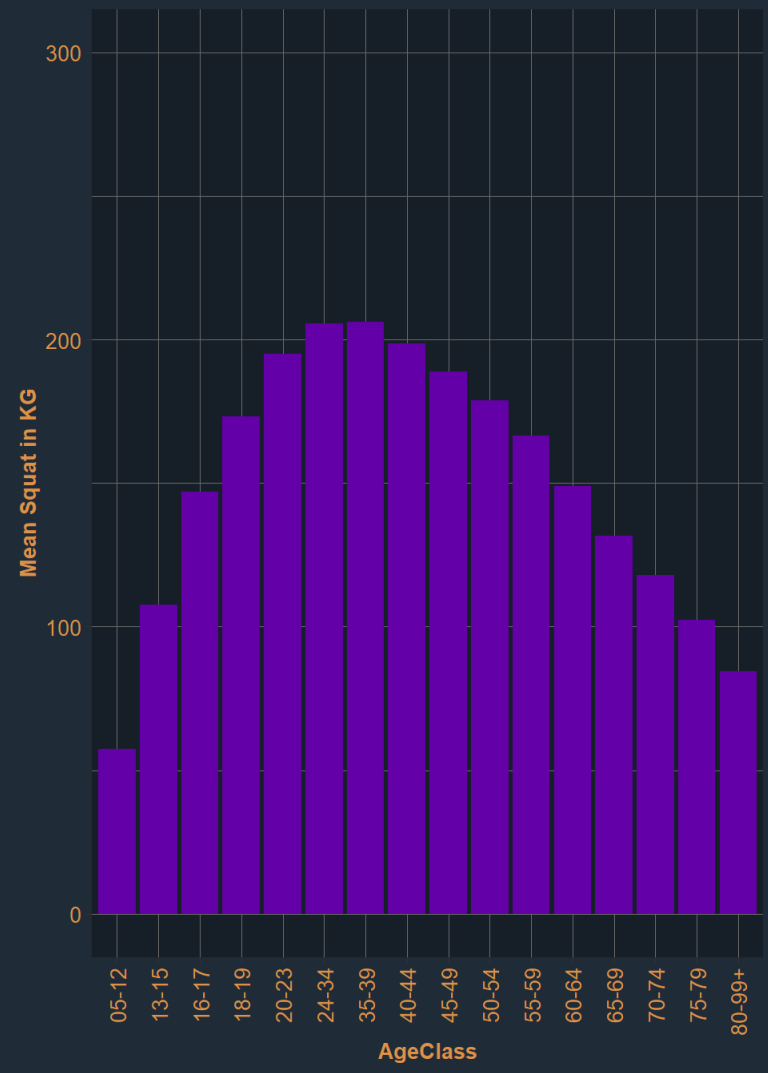
As an interim conclusion, we can note that while bodyweight is influential for both male and female performances in all lifts, it is so to a varying degree, accounting for significant portions of the variance in the total (36% for the men and 20 for the women), squat (34/19), bench press (37/14) and deadlift (28/17) performances. There are a few peculiarities, however, as we can observe a relatively low determination coefficient in the women's bench press and the male deadlift as noticeable downward outliers.

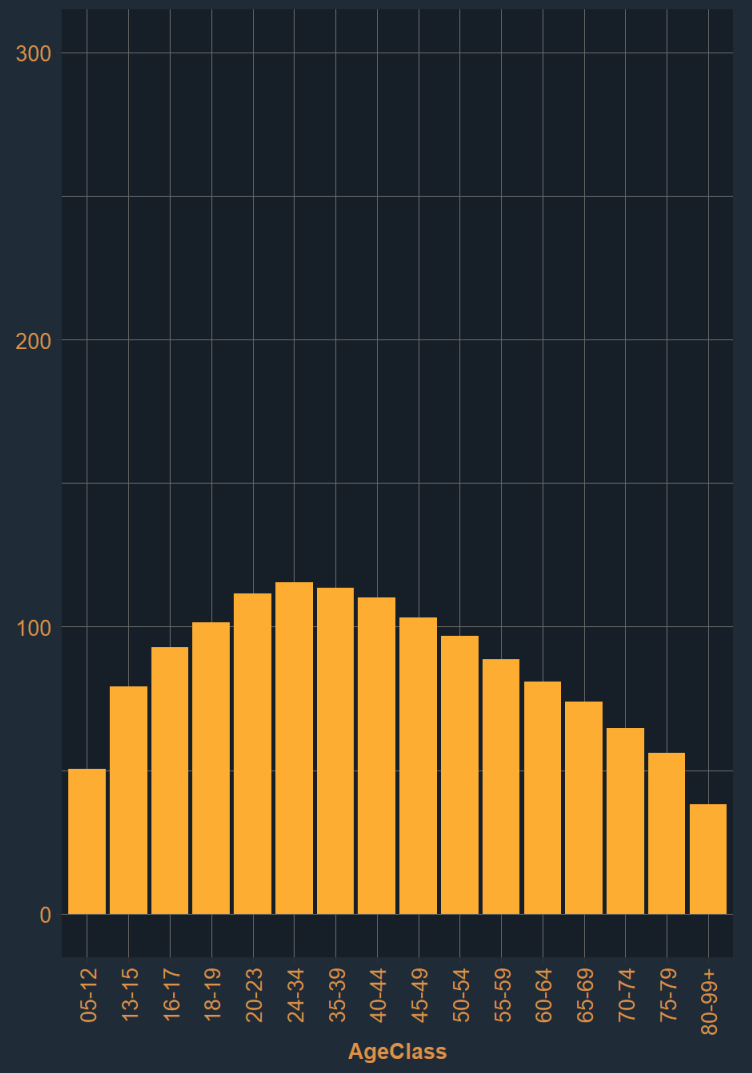
Influence of Age

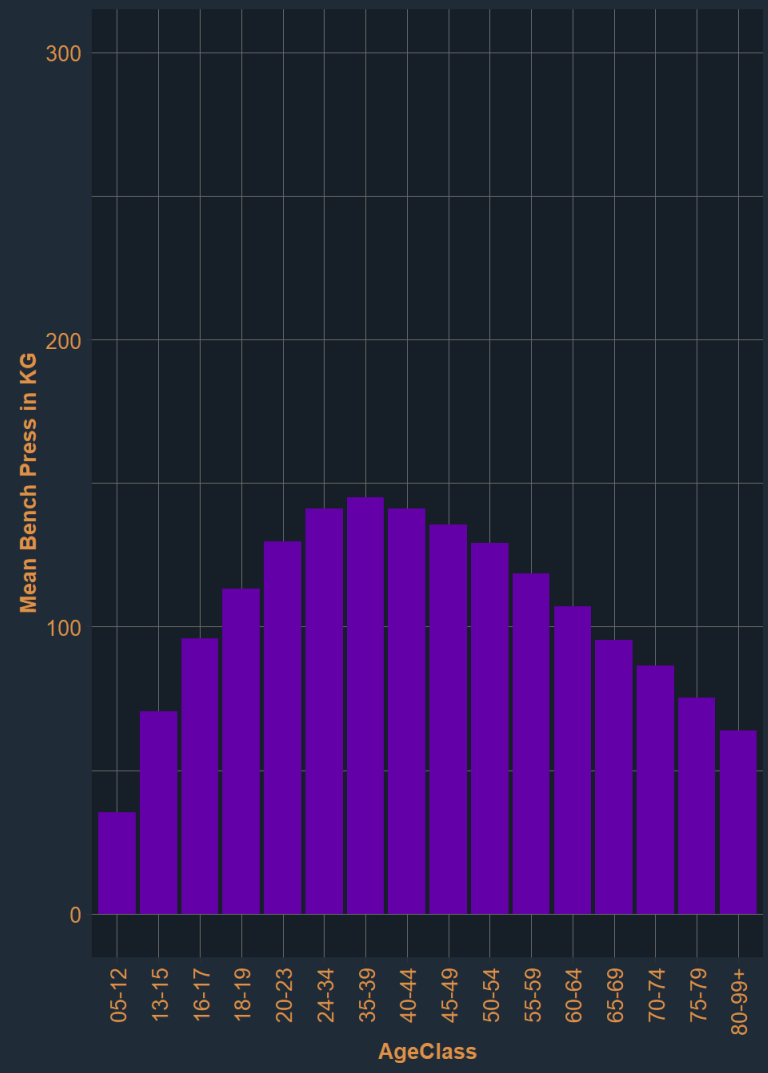
So, now that we know a little more about the influence of bodyweight, let us explore the impact of age, another factor which we can assume to be important right from the get-go. This will, however, require a different approach as the one I took with bodyweight, since a simple linear regression will not be a good fit for the data. Age and performance do not stand in a linear relationship to each other: Athletes only increase their performances to a certain point of age, after which a decline can be expected. A simple barplot of mean performances across age groups can indicate this nicely:

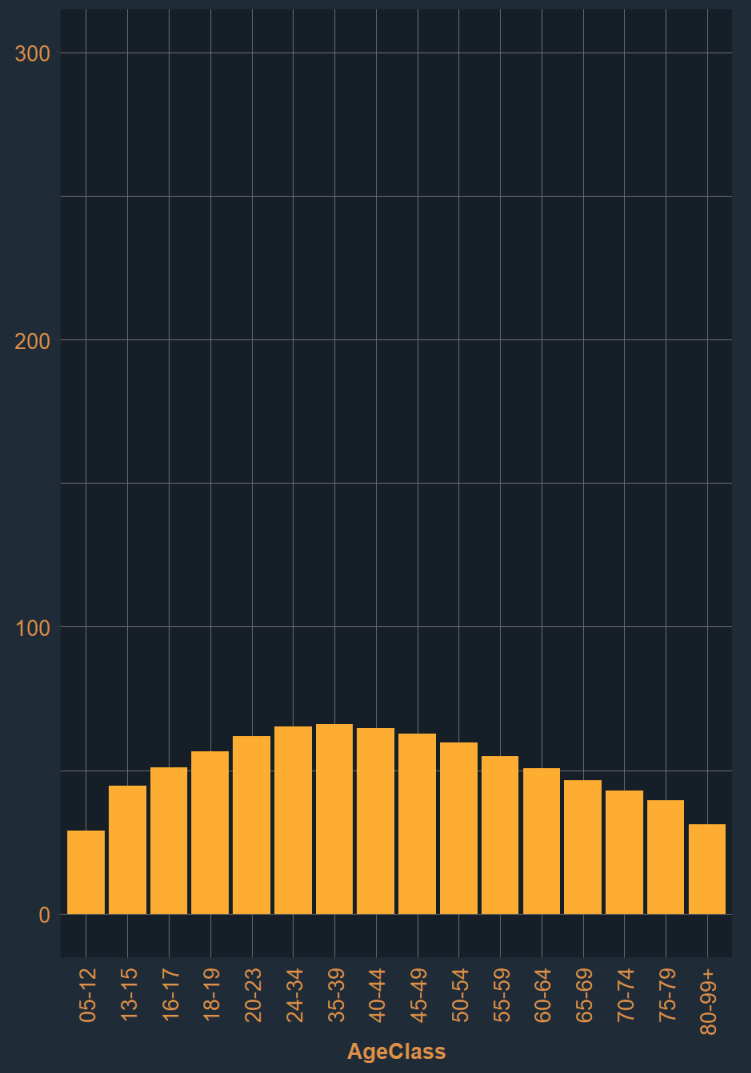


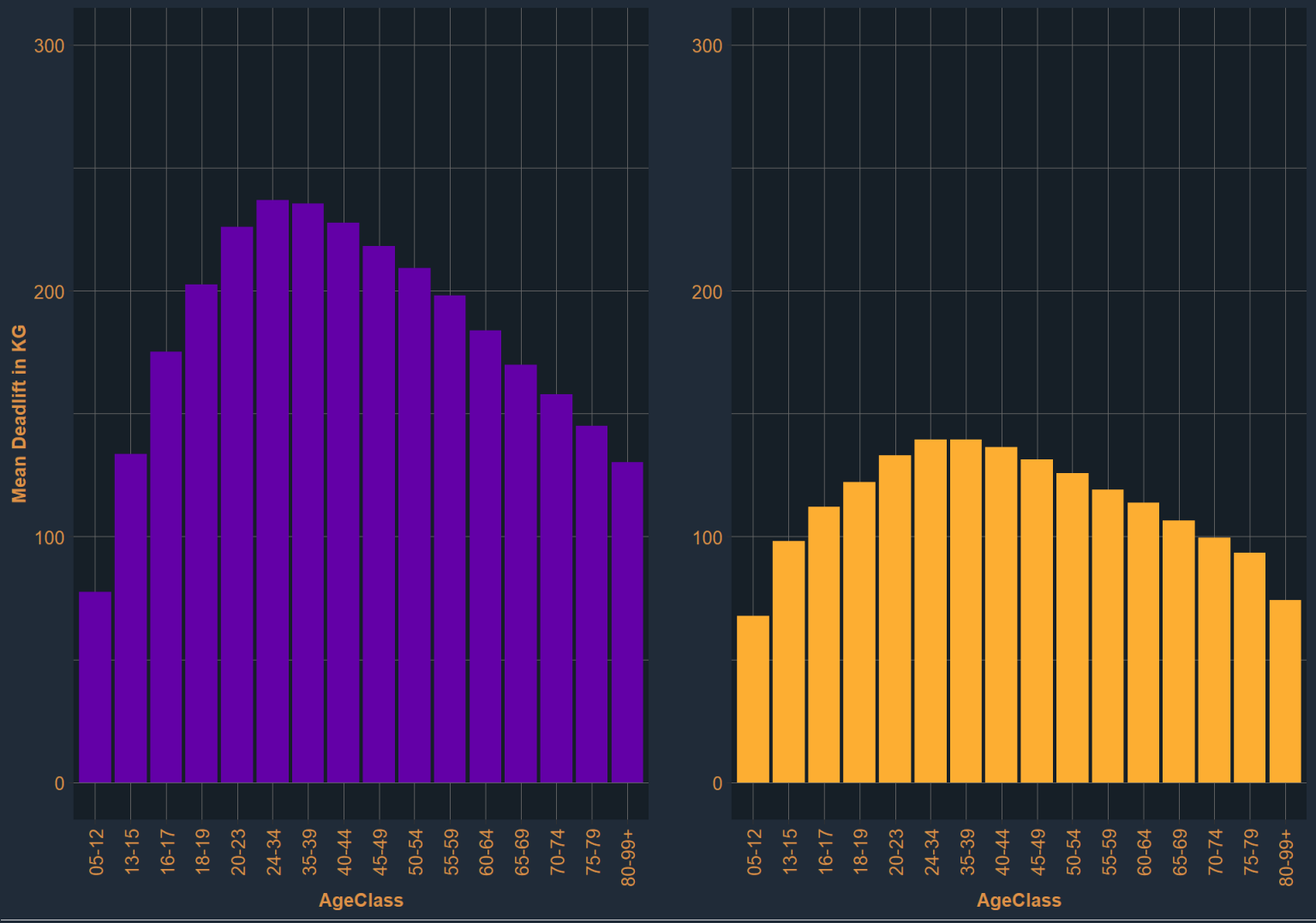
The trend here is as clear as it is intuitive: We can expect the highest performance in the medium age groups between 24 and 39 years of age. This holds true for both men and women and applies, with slight variations, to the total, squat, bench press and deadlift alike:


300
200
100
0
Mean Squat in KG
05-12
13-15
16-17
18-19
20-23
24-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80-99+
AgeClass

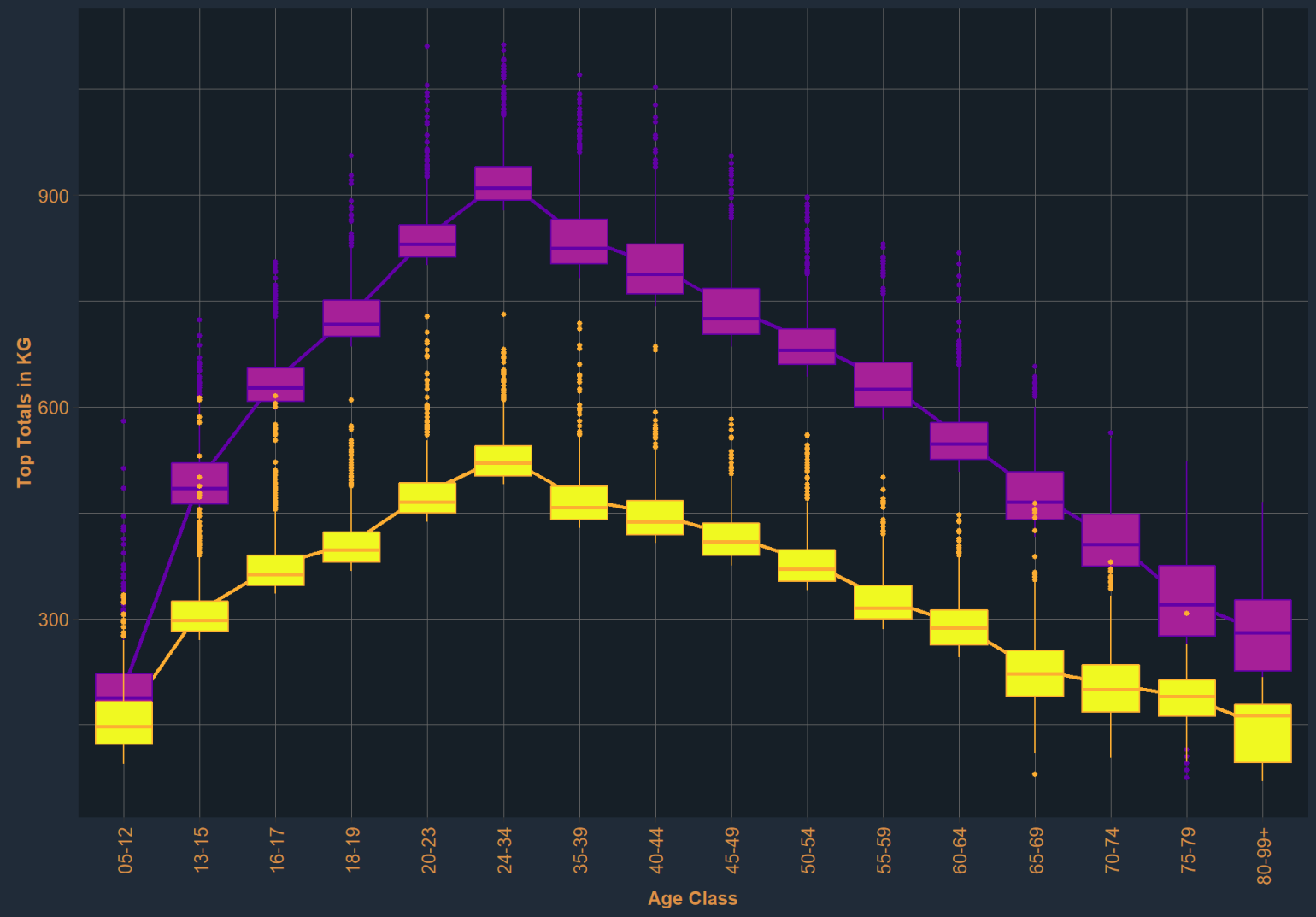

300
200
100
0
05-12
13-15
16-17
18-19
20-23
24-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80-99+
AgeClass


300
200
100
0
Mean Bench Press in KG
05-12
13-15
16-17
18-19
20-23
24-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80-99+
AgeClass


300
200
100
0
05-12
13-15
16-17
18-19
20-23
24-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80-99+
AgeClass



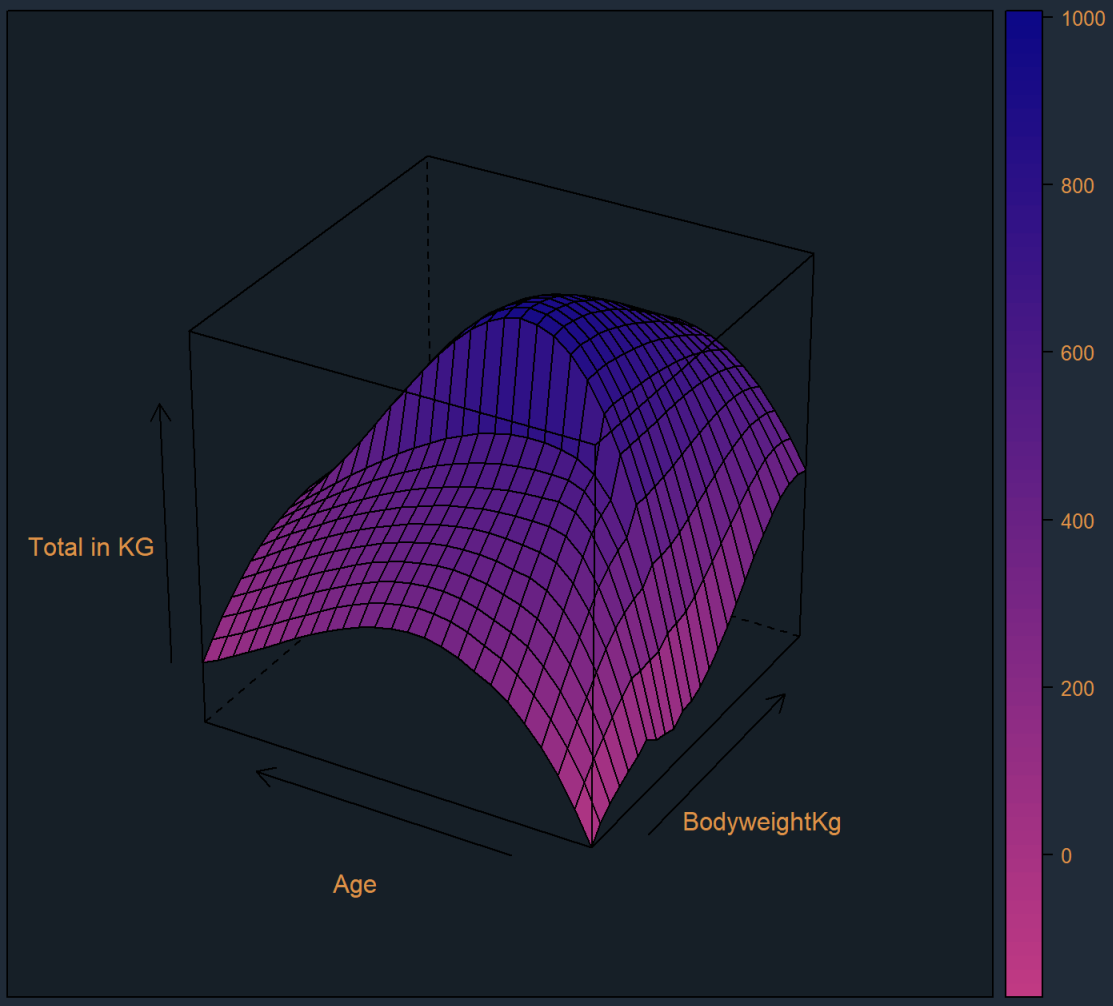
While this is certainly an expected result, it still leaves a somewhat bitter taste. After all, being reminded of an inevitable downfall in performance is not something too encouraging when taking a look at something as taxing as competitive powerlifting. And I do not want to accidentally discourage anyone from further following this path, so instead of taking a look at what it typical, let us take a look at what is possible: What are the top totals in all our age classes?



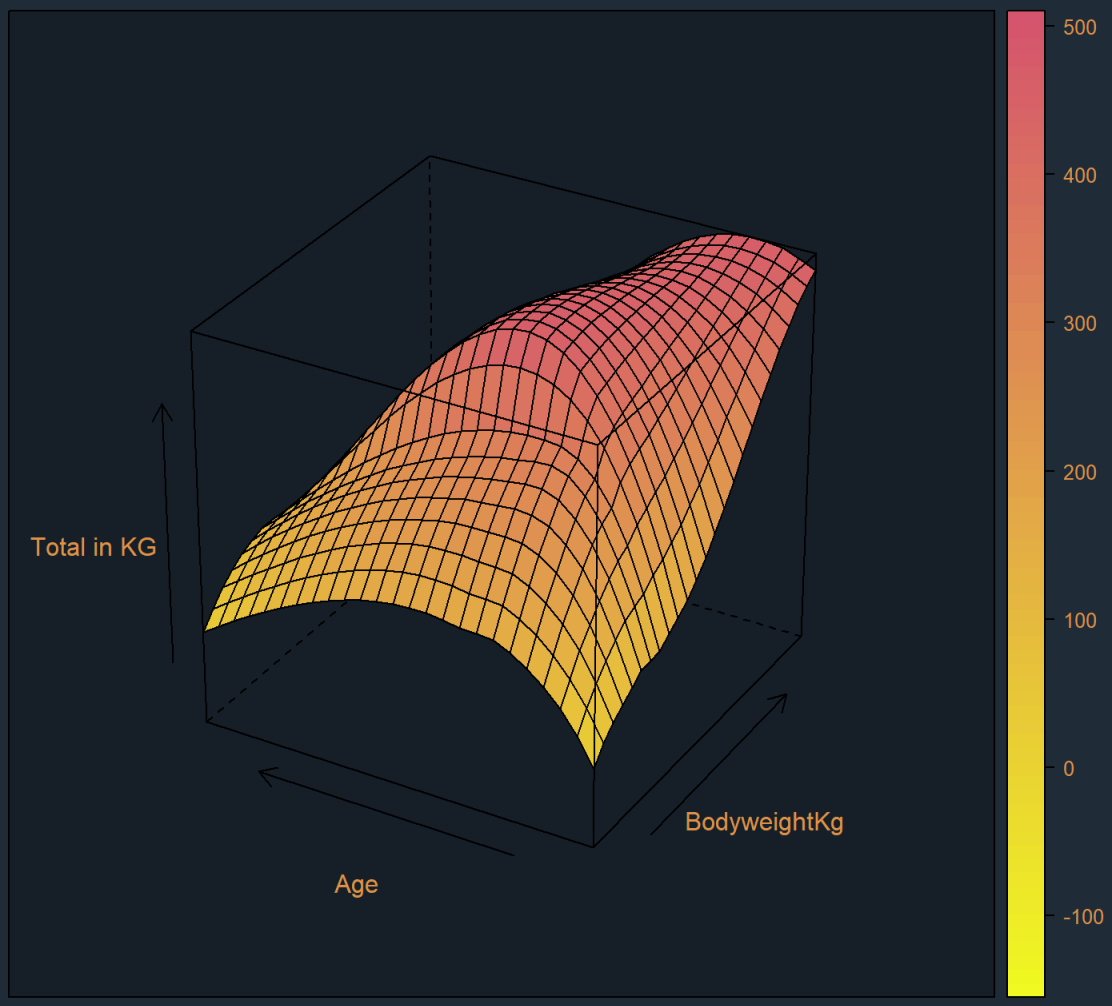
This, at least in my opinion, manages to take the edge of slightly. While we can still observe the overall decline in performance across age classes after the peak is reached between 24 and 34, there are still quite a number of incredibly impressive performances even in the higher age classes. Some men manage crossing 600, some women 450KG totals while competing in the age class of 65 to 69 year-olds, for example. Incredibly impressive stuff and certainly something to aspire toward!

Bringing Both together - Age and Bodyweight

Okay, now that the morale is high again, let us get back to the core topic here: Clearly both age and weight are influential factors in an athlete's performance. So, can we visualize the interplay of both? To do this, I will make use of the lattice package and sample from the original dataset, since creating this plot requires quite a bit computation and using the entire dataset would take an unnecessarily long time. Please note that this plot is a little more complex than what we have seen so far. Instead of using only the data we analyzed so far, we calculate smoothed values and try to predict performances over different age and weight groups based on our sampled data. This plot is intended to show the interplay of age and bodyweight *in general* and should not be taken at face value for any particular case.



Still, we are able to take away a few insights from this. First, we can identify both the trends of medium age and higher bodyweight being beneficial for total performance. There seems to be, however, a point in bodyweight after which the general trend of higher bodyweight equaling higher total reverses and additional bodyweight becomes more of a detriment. A third aspect is the bump in the middle area of the wireframe, which might be explainable by a high density of very strong lifters in the particular area of medium age and high, but not overly high bodyweight, causing our model to predict particularly high totals in that specific area.



If we construct a similar plot for the women, the general pattern can still be observed: Bodyweight has a positive influence on the total up to a certain point, as does age. Interestingly enough, the same bump in the middle of the wireframe can be observed for the women as well, even though noticeably smoother than the one we discussed for the male model. Another peculiarity we can notice in this plot is the prediction of very high totals for lifters at very high bodyweight and very young age. This seems rather counter-intuitive and I believe it to be caused by a lack of data in that specific category, causing our model to make inaccurate predictions as a result of faulty extrapolations.

Statistical Testing

Tested vs. Untested Competition

Now that we understand the demographics of the competitors as well as a variety of meaningful factors in their performance a little better, I will move onto the second part of this analysis: Testing.

As a first topic worth investigating: We know which athletes compete tested and which compete untested. I want to explore what size of difference one can expect when deciding for competition in either bracket. To do this, I will calculate a t-test which determines whether there is a statistically significant difference in the performance between both groups or not. This test makes assumptions about the data, which I will test in the code, but not include in the

output, since doing so would take make this section overly technical. If you are interested nonetheless, feel free to check out the comments added to the code chunks in the Markdown document!

I decided against sampling the data, since I am more interested in how big the actual difference in the entire populations of athletes will be. Doing this will almost guarantee a significant difference between both samples, since the size of the dataset is still massive and a huge sample size constitutes one important factor in decreasing the threshold for potential differences to be considered statistically significant.
This means we can almost certainly expect a significant difference to turn up. This means we are at this point more interested in whether or not the differences we observe here would turn out to be meaningful in practice.

I believe that we should, however, not expect the differences in the results to be all that impressive, since being in the 'Tested' category does not mean that the lifter in question was actually subject to a drug test. Instead, this column only indicated that the lifter delivered this performance at an event at which drug testing took place.

Still, let us take a look at the results:

```
##
##  Welch Two Sample t-test
##
## data:  tested_m$TotalKg and untested_m$TotalKg
## t = -72.41, df = 113326, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99.9 percent confidence interval:
##  -39.72315 -36.26974
## sample estimates:
## mean of x mean of y
##  526.1244  564.1208
```

Completely contrary to my intuition, we can actually observe a massive difference of somewhere between 36.27 and 39.72KG in the men's total. The untested population clocks in with a total that is not only (as expected) statistically significant from the tested population's but massively so, with the difference amounting to a little more than 7% of the tested total!

In hope of shedding light on how exactly these differences come to be, let us explore the existing differences in the three main lifts by applying the same test. First off, the squat:

```
##
##  Welch Two Sample t-test
##
## data:  tested_m$Best3SquatKg and untested_m$Best3SquatKg
## t = -52.075, df = 112167, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -12.24972 -11.09499
## sample estimates:
## mean of x mean of y
##  184.8072  196.4796
```

Secondly, the bench press:

```
##
##  Welch Two Sample t-test
##
## data:  tested_m$Best3BenchKg and untested_m$Best3BenchKg
## t = -65.502, df = 106903, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99.9 percent confidence interval:
##  -11.010633  -9.957275
## sample estimates:
## mean of x mean of y
##  125.2702  135.7542
```

And finally, the deadlift:

```
##
##  Welch Two Sample t-test
##
## data:  tested_m$Best3DeadliftKg and untested_m$Best3DeadliftKg
## t = -59.966, df = 112844, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99.9 percent confidence interval:
##  -13.45410 -12.05434
## sample estimates:
## mean of x mean of y
##  215.1990  227.9532
```

Somewhat in accordance with the differences we observed between the totals of tested and untested male lifters, we can notice differences of roughly 11-12Kg in the squat, 10-11Kg in the bench press and 12-13KG in the deadlift. While these are no differences that cannot be overcome by diet, lifestyle or wisely planned and disciplined training, they are certainly worthy to take note of.

Let us take a look at how the women compare to this. First off, the total between tested and untested female lifters:

```
##
##  Welch Two Sample t-test
##
## data:  tested_f$TotalKg and untested_f$TotalKg
## t = -19.74, df = 51251, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99.9 percent confidence interval:
##  -10.219984  -7.299427
## sample estimates:
## mean of x mean of y
##  298.3817  307.1414
```

The differences in the women’s total are far less pronounced than those in the male population, residing at approximately 8.75KG, which would be around 2.9% of the untested

total: Less than half of what we observed in the male population! Let us also take a look at the differences in the three main lifts in order to better understand what is going on here.

As usual, we will start with the squat:

```
##
##  Welch Two Sample t-test
##
## data:  tested_f$Best3SquatKg and untested_f$Best3SquatKg
## t = -9.7895, df = 50915, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99.9 percent confidence interval:
##  -2.542326 -1.263134
## sample estimates:
## mean of x mean of y
##  106.8137  108.7164
```

Then we can take a look at tested and untested bench press...

```
##
##  Welch Two Sample t-test
##
## data:  tested_f$Best3BenchKg and untested_f$Best3BenchKg
## t = -11.219, df = 49185, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99.9 percent confidence interval:
##  -1.6528063 -0.9030954
## sample estimates:
## mean of x mean of y
##  60.85759  62.13554
```

...as well as the deadlifts.

```
##
##  Welch Two Sample t-test
##
## data:  tested_f$Best3DeadliftKg and untested_f$Best3DeadliftKg
## t = -22.142, df = 52305, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99.9 percent confidence interval:
##  -4.662760 -3.456142
## sample estimates:
## mean of x mean of y
##  130.2906  134.3500
```

The difference between tested and untested population for the women is around 1-2.5Kg for the squat, 0.9-1.6Kg for the bench press and 3.4 to 4.7KG for the deadlift, which seem to be in accordance with what the differences observed for the totals.

While these differences seem pretty substantial, especially in the men's population, I hesitate to interpret them at face value. As previously mentioned, the only information we have available is whether or not the results count as drug tested. There is no information about

whether a specific lifter was drug-tested or not. Instead, we compare the lifters who compete at tested to those who compete at untested events. While there might very well exist a meaningful overlap, there is also a significant chance of other factors coming into the equation here. After all, we did not check and control other factors, like bodyweight or age in both groups. Still, the differences we can observe here seem substantial enough to be considered meaningful in practice: Whatever the exact composition of causes for the differences might be, the competition at untested meets is noticeably steeper than that at tested meets, an insight that should be taken into consideration when planning for competition.

Athletes Competing at Home vs. Athletes Competing Abroad

Another topic we can take a look at using the Standard t-test is whether or not there is a statistically significant difference between athletes competing in their home country and athletes competing abroad. In general, I want to test two of my hypothesis against another: Number one being that travel, acclimatization and the competition in a less familiar environment will cost an athlete strength, resulting in a lower total. Against that, there is also the possibility that only a certain type of athlete will even compete abroad in the first place - those who have already reached a sufficiently high level in strength, causing them to seek competition on a higher level, which might very well take place in another country.

Let us take a look!

```
##
##  Welch Two Sample t-test
##
## data:  competing_home_m$TotalKg and competing_abroad_m$TotalKg
## t = 84.579, df = 234563, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  35.70536 37.94850
## sample estimates:
## mean of x mean of y
##  545.5556  508.7286
```

Well, this result speaks for hypothesis number one and in a surprisingly clear fashion as well. The differences we can observe here are similar to those resulting from the comparison between tested and untested lifters above. Let us take a look at whether a similar phenomenon can be observed in the women’s population too.

```
##
##  Welch Two Sample t-test
##
## data:  competing_home_f$TotalKg and competing_abroad_f$TotalKg
## t = 55.662, df = 97341, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  19.62360 21.52798
## sample estimates:
```

```
## mean of x mean of y
##  306.0294  285.4536
```

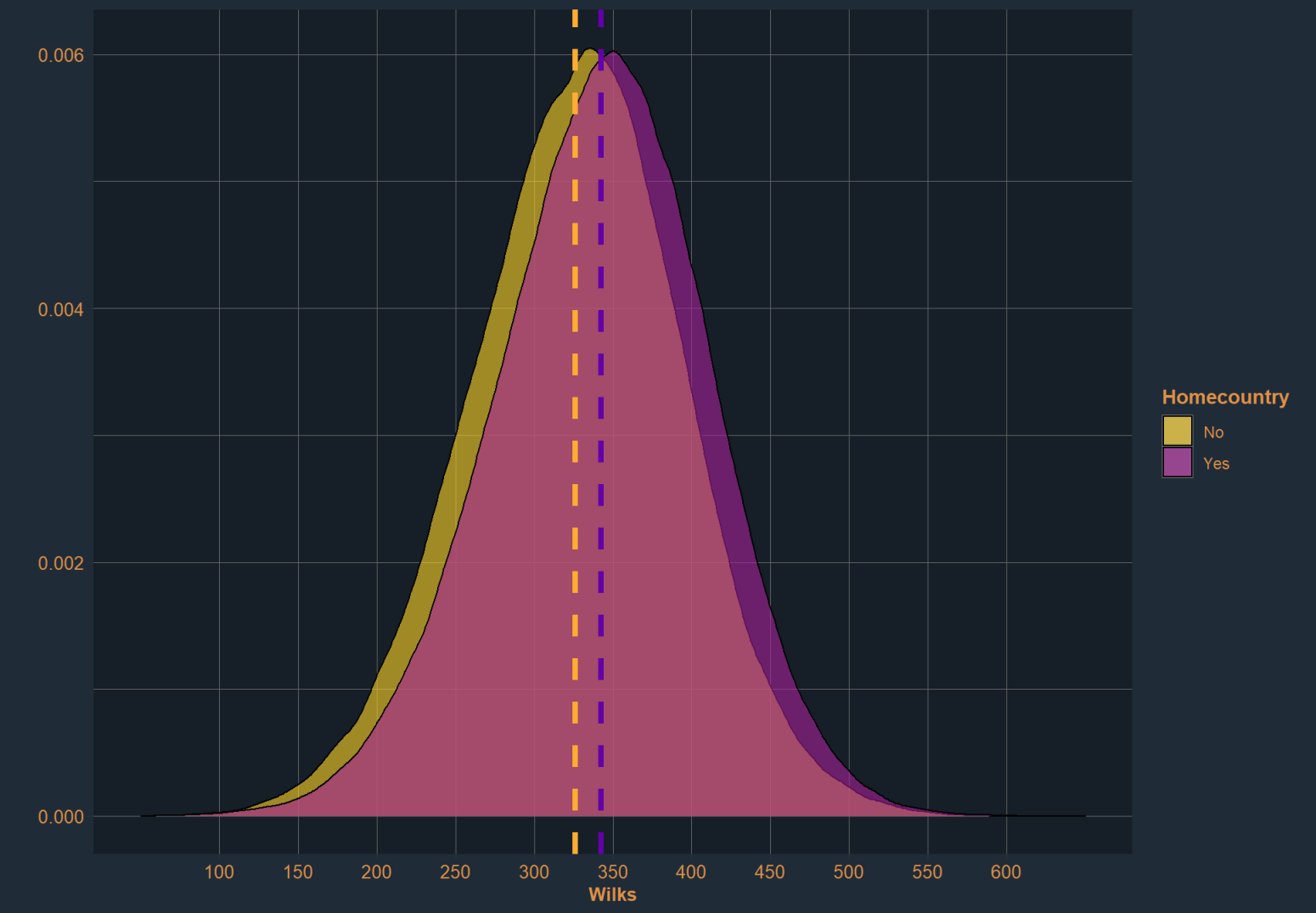
Same thing for the women, higher total for those competing at home. In this case, the differences we can observe are even more pronounced than those between the tested and untested women - Which is certainly a surprise.

I am interested in what we are observing here. These differences seem to be *too* large to be cause only by the factors I hypothesized about. One way to check what is going on is to look at stronger athletes:

```
##
##  Welch Two Sample t-test
##
## data:  competing_home_m$TotalKg and competing_abroad_m$TotalKg
## t = 34.314, df = 95724, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  14.32801 16.65374
## sample estimates:
## mean of x mean of y
##  621.5502  606.0593
```

```
##
##  Welch Two Sample t-test
##
## data:  competing_home_f$TotalKg and competing_abroad_f$TotalKg
## t = 5.7252, df = 11041, p-value = 1.06e-08
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  2.621835 6.911896
## sample estimates:
## mean of x mean of y
##  390.6825  385.9157
```

As we can see, the difference between athletes competing at home and those that compete abroad is far smaller in the group of stronger athletes. Still, there actually seem to be meaningful differences between people competing at home and those competing abroad across the entire spectrum of wilks score:



To be frank, I am taken aback by these findings and am unsure of what an adequate interpretation would look like. A difference of several points in wilks score across the entire spectrum of strength between both groups of athletes seems almost to good to be true. I am simply going to use the same escape I udes in the interpretation of the differences between tested and untested athletes: There might be systematic differences between both groups that I did not notice in my background checks and I would be careful to interpret these results at face value. Still, even if the differences we observe here are only partially caused by the location the athletes compete in, it might be a meaningful influence and hence worth considering when planning for competition.

Differences Between Classes of Equipment

For the next part of the analysis, I will again pick up the topic of equipment used in competition. I have already tried to showcase how important the various forms of equipment are for certain lifts by displaying the difference between the best bench press performances in a few categories. However, I would like to quantify these differences in the same way I tried to do in respect to drug testing and location of competition just now: What difference can we expect for the different classes of equipment? To do this, I will have to calculate an ANOVA, which, like the t-test, has certain assumptions about the data, for which I will check in the code. Again, if you are interested, feel free to take a look. A more important point of note, however, is that in order to calculate the ANOVA or equivalent tests, I have to sample from the data. This means that the results we obtain should be taken with a grain of salt, since they are

obtained using only a small portion of the data. When analyzing another sample, they might very well vary.

Sidenote and slight spoiler: The data only partially fulfills the prerequisites of the ANOVA, which means I resorted to calculating a Kruskall-Wallis test instead, which is roughly equivalent to the ANOVA, but better suited to handle data like ours, which suffers from heteroscedasticity or unequal variance across groups. So if I am talking about the results of the Kruskall-Wallis test from now on and you are wondering what happened to the planned ANOVA: We simply exchanged one test for another one, which is better suited for the data while serving the same purpose.

Let us take a look at the results:

```
##
##  Kruskal-Wallis rank sum test
##
## data:  anova_data_m$TotalKg by anova_data_m$Equipment
## Kruskal-Wallis chi-squared = 431.51, df = 4, p-value < 2.2e-16
```

The output of the Kruskall-Wallis-Test itself is, unfortunately, relatively boring in this case, only indicating that statistically significant differences between groups have been found. In order to find out between which groups these differences exist, we have to dig a little deeper, using pairwise t-tests:

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  anova_data_m$TotalKg and anova_data_m$Equipment
##
##            Multi-ply Raw     Single-ply Unlimited
## Raw        < 2e-16   -       -          -
## Single-ply < 2e-16   0.0635  -          -
## Unlimited  0.0084    < 2e-16 < 2e-16    -
## Wraps      < 2e-16   4.5e-05 6.1e-09    < 2e-16
##
## P value adjustment method: bonferroni
```

Significant results across the board!

Differences can be found between all combinations of groups, except for the comparison between Raw and Single-Ply. This is certainly surprising! Notice however, that the p-value for the significance test between Raw and Single-Ply is close to 5%, our threshold for significant differences. We also resorted to the Bonferroni correction, which is very conservative and, as a consequence, might have just tipped our results over the edge of not detecting a significant difference anymore.

Let us move to the women's data and then I will try to visualize these results, which should shed a little more light on the magnitude of the differences we can expect in competition.

```
##
##  Kruskal-Wallis rank sum test
```

```
##
## data:  anova_data_f$TotalKg by anova_data_f$Equipment
## Kruskal-Wallis chi-squared = 167.82, df = 3, p-value < 2.2e-16
```

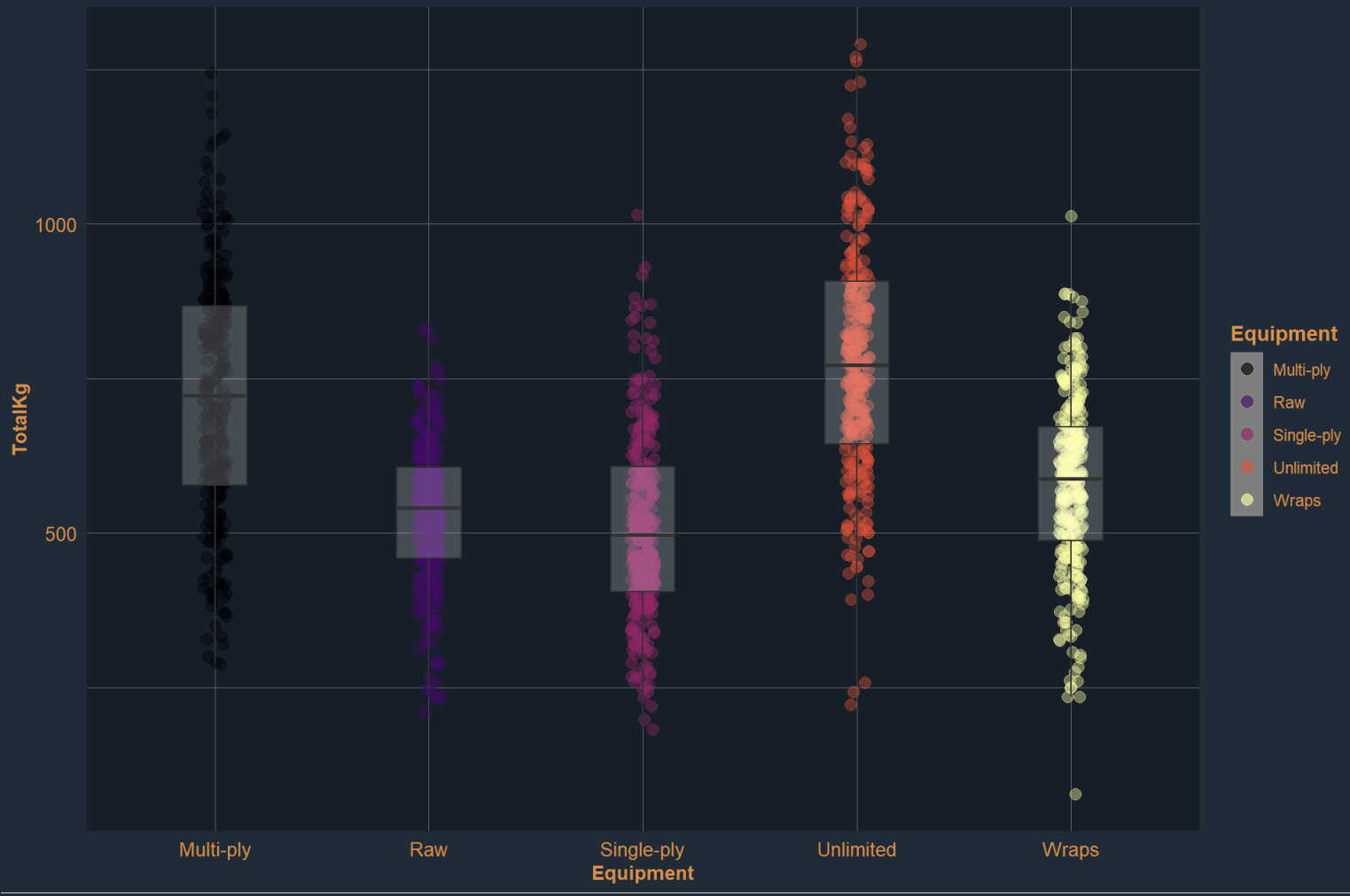
As it was the case for the men, there are significant differences between the women's groups as well. Not particularly surprising, but nice to see.

Let us take a look at where these differences can be found:

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  anova_data_f$TotalKg and anova_data_f$Equipment
##
##            Multi-ply Raw  Single-ply
## Raw        <2e-16    -    -
## Single-ply <2e-16    1.00 -
## Wraps      <2e-16    0.47 0.57
##
## P value adjustment method: bonferroni
```

While the comparisons between Multi-Ply and the other groups all indicate significant differences, none of the other comparisons do! This at the very least indicates how impactful Multi-Ply is. However, we should be careful when interpreting these results, since the bonferroni correction we have used is very conservative, causing our test to only pick up sufficiently large differences between groups, while discarding smaller ones. I strongly suspect this to be the case here.

To shed a little more light on this issue, I would like to visualize the results of our Kruskall-Wallis-Test, which should help to get a more intuitive understanding of the differences in total we are looking at here. First off, the data of the men's groups:



These results of our male samples are clearly indicating differences between all classes of Equipment. Interestingly enough, the lifters in the Single-Ply class perform slightly worse than those in the raw category in this sample, while Multi-Ply and especially Unlimited are clearly above the rest in terms of performance. Single-Ply delivering the lowest performance of the bunch seems like a particularity of this sample and is highly unlikely to be the case in the entire population. Disregarding this particular result, the advantage of Multi-Ply and Unlimited equipment is noticeable and delivers an estimation of what difference can be expected between these classes.

Next up: The same visualization for the results of the women’s test:

As we were already able to see in the results of the Kruskall-Wallis-Test of the women's data, there is only one class of equipment, which really stands out: Multi-Ply, which sports a total only slightly below 400KG. All other classes of equipment (those for which there was enough data to include them in the evaluation, hence no 'Unlimited' in this plot) are around the 300KG mark.

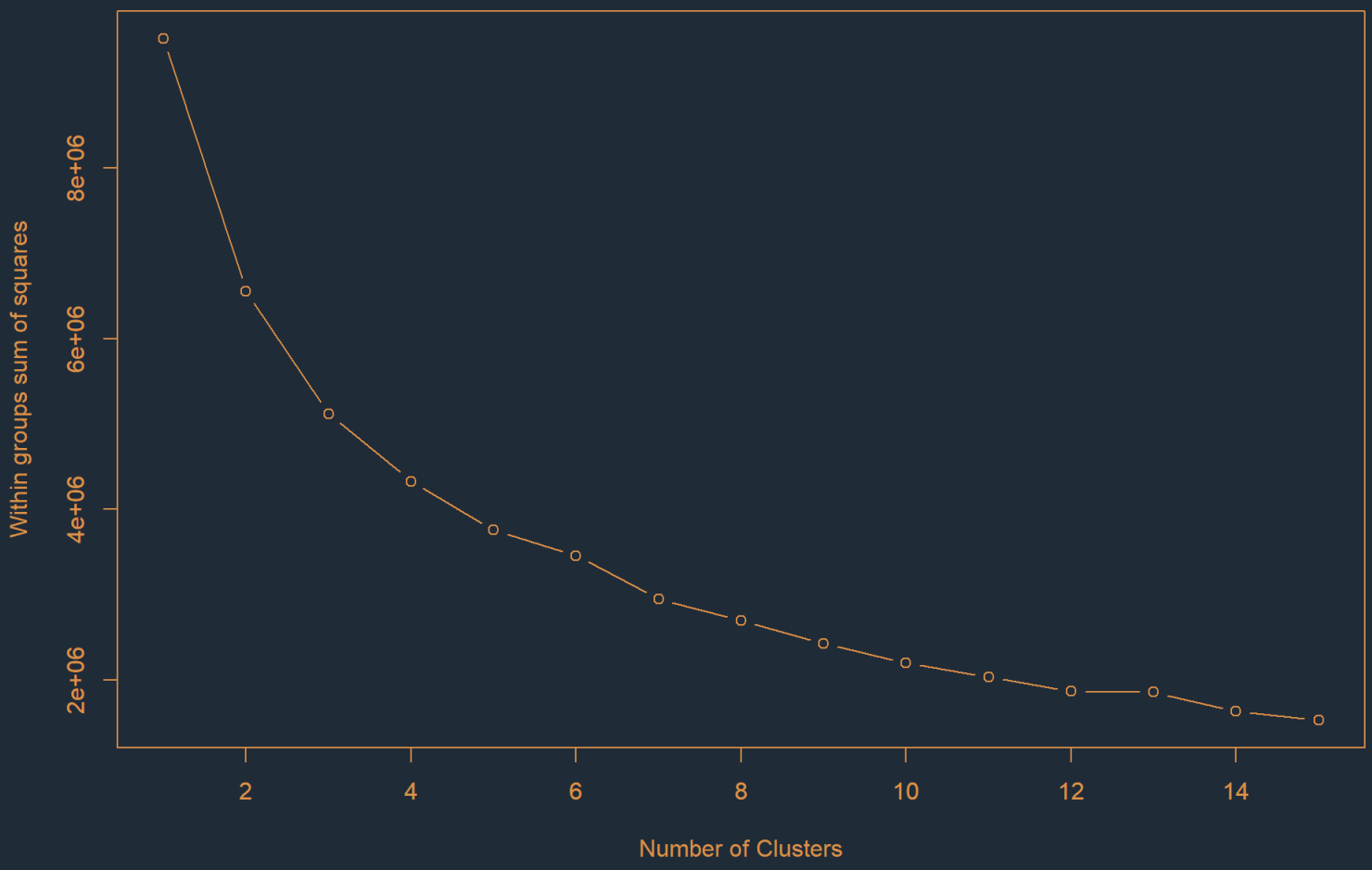
Clustering - Are There Different Types of Lifters?

Now for something close to my heart. I am interested in whether or not we can use the data of the open powerlifting project to find different clusters of lifters. The idea here is that people with different body types differ systematically in the way they obtain their total: Longer arms, for example, might be greatly beneficial when deadlifting, but less so when bench pressing. Initially, I wanted to explore lifters height in exploring potential relationships, but in the meantime managed to discover that the data does not contain any measure of height. Whelp.

As a second approach, I then decided to explore the percentages of squat, bench press and deadlift in each lifters total: Can we make out certain groups of lifters that are very similar in terms of the percentages of squat, bench and deadlift that constitute their total, while being

different from lifters in other groups? Are there meaningful differences between these groups in terms of other characteristics like age or the lifters overall strength?

To get this done, I will make use of a clustering algorithm called *k-means*, which will group the data into k clusters, for each of which there will be one prototype or *centroid*, best suited to represent that particular cluster. In order to figure out how many clusters exactly we can or should group the data into, it is wise to consider both the data's structure and the theoretical framework under which we want to evaluate it. Let us start by taking a look at an elbow plot, which will be able to provide some orientation as to how many clusters make sense:




```
##  [1] 9518701 6551406 5116591 4327144 3753350 3456046 2946145 2695668 2425179
## [10] 2194945 2033343 1862803 1854860 1629944 1527078
```

This plot illustrates how many clusters we would need to explain a majority of the variance in the data. It makes sense to go with a number of clusters after which the additional amount of explained variance begins to flatten significantly (i.e. the 'elbow' of the curve). In this case, I would argue for a number of four clusters into which we can partition our lifters: Using this amount enables us to explain a great deal of the variance in the data while only resorting to a moderate amount of clusters, which, in turn, keeps the results interpretable without being too confusing. Still, other interpretations are certainly viable and any number between three and eight clusters could make sense, depending on what hypothesis one might want to explore.

If continuing along the road of four clusters, we can then create the clusters and display each cluster's centroid, the single data point that best represents the respective cluster. In our case of four clusters (and therefore centroids), we would end up with the following results for the men's category:

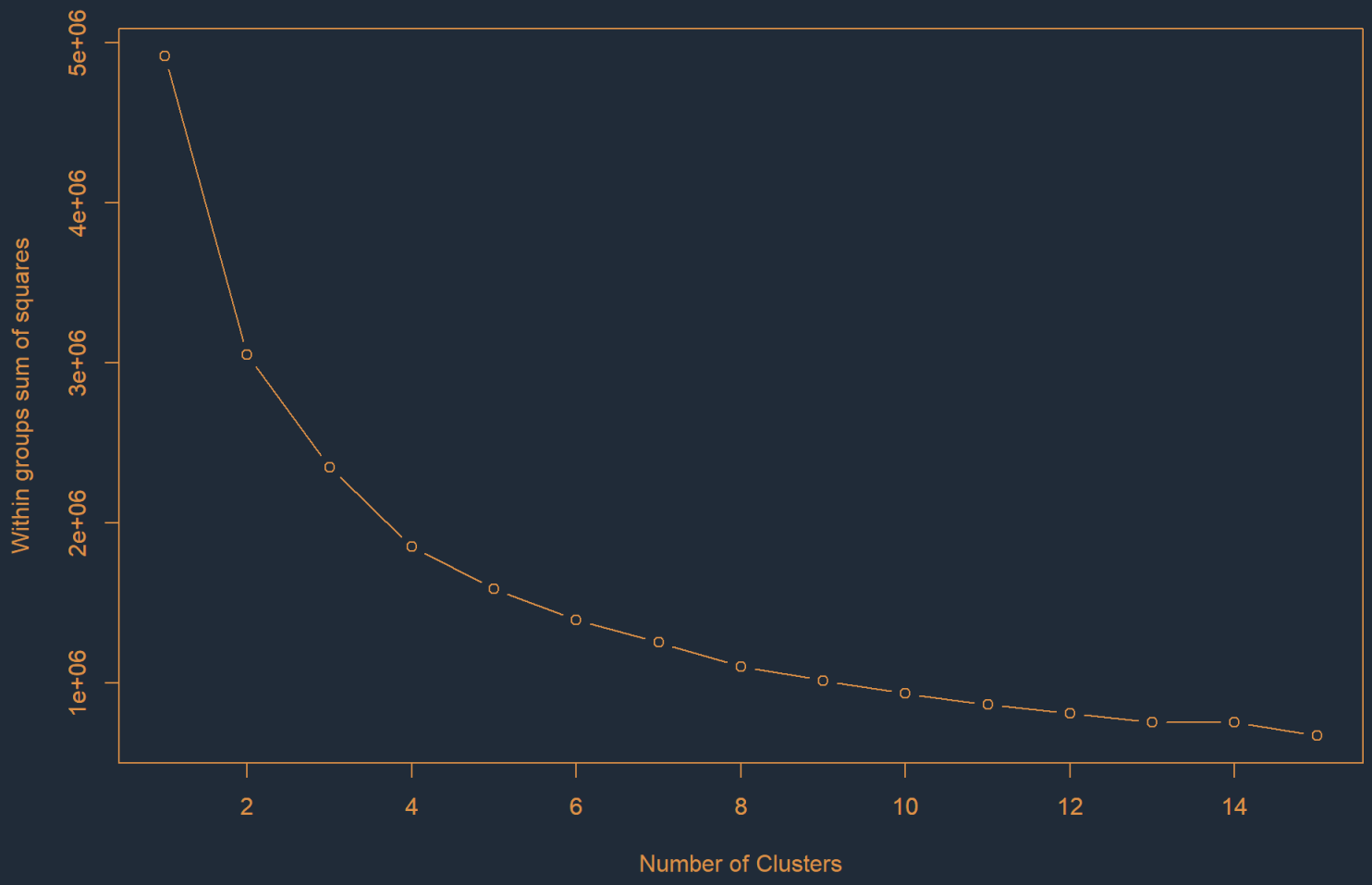
```
##   sq_percentage b_percentage dl_percentage
## 1      32.89216     26.96042      40.14753
## 2      37.19632     25.00167      37.80208
## 3      36.12855     22.12547      41.74605
## 4      32.30454     22.43132      45.26422
```

There are a few things to take note of in these results. First of all, we can notice clear distinctions between our clusters: Two clusters, 1 and 4, obtain a relatively low percentage of their total from the squat at around 32-33%. The two other clusters, 2 and 3, clock in with a noticeably higher squat percentage of 36-37%. When looking at the percentages these groups obtain in the bench press, a similar pattern, albeit with a slight twist, can be observed: Both our groups of high-percentage and low-percentage squatters can be further divided into a group of high (25-26% of the total) and low-percentage benchers (22% of the total). This means we end up with a High-Squat/High-Bench, a High-S/Low-B, a Low-S/High-B and a Low-S/Low-B group. Since there is only 100% of the total to distribute across the three lifts, this also dictates what the final row, containing the deadlift percentages, will look like: Our High-S/High-B group grabs the lowest percentage of the total (37%) from the deadlift, while the Low-S/Low-B group obtains the highest percentage of 45%. Both our Low-S/High-B and High-S/Low-B groups are around the middle mark with 40 and 41 percent, respectively.

An interesting observation we can make here is that while our clusters combine high squat percentages with both high bench and high deadlift percentages, a notable absentee is the combination of high bench and high deadlift percentages: Groups three and four, who both obtain a relatively low percentage of their total from their bench, get significantly more percentage out of their deadlift than the other groups. An intuitive reason for this peculiarity might lie in anthropometric properties of the competitors, like for example their arm length, which might be disadvantageous when bench pressing but helpful when deadlifting.

Unfortunately, the dataset does not provide any information to further explore this hypothesis. I would have, for example, liked to explore the average height of athletes in all clusters, which might have delivered further insights. We also do not get any information about whether the deadlifts were performed in conventional or sumo stance, which could have been another interesting aspect to look into. Still, these results provide adequate grounds for further exploration and I believe that it makes sense to take a look at the relation each cluster has with other metrics we *do* have access to: Age, total performance, wilks and bodyweight chief among them.

But before that, we will explore the women's data, starting, as we did above, with the elbow plot to help us choose a number of clusters.




```
##  [1] 4916617.3 3049390.2 2348732.0 1853121.0 1588242.2 1394742.9 1254959.7
##  [8] 1103672.8 1017435.6  936455.6  865589.3  812200.4  754149.4  757084.4
## [15]  673066.7
```

As with the male population, arguments can be made for anything between three and six (or even more) clusters. However, four clusters seems to make the most sense when following the elbow plot, while also having served our purpose well during the analysis of the men's data. So, let us go with four clusters again.

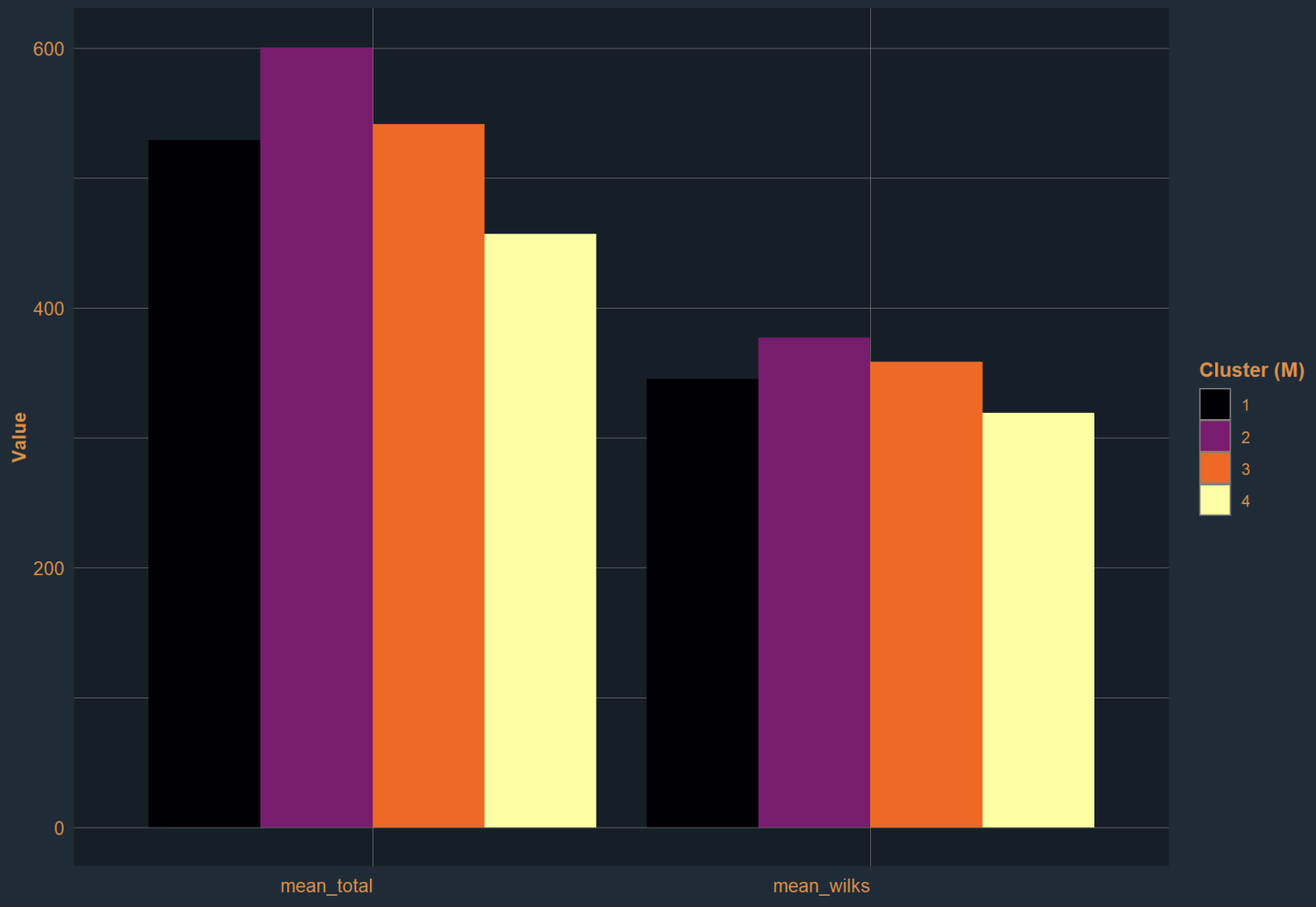
```
##   sq_percentage b_percentage dl_percentage
## 1      34.47981     23.35480      42.16545
## 2      35.93274     18.73114      45.33617
## 3      31.01650     20.28679      48.69682
## 4      38.89447     20.16081      40.94480
```

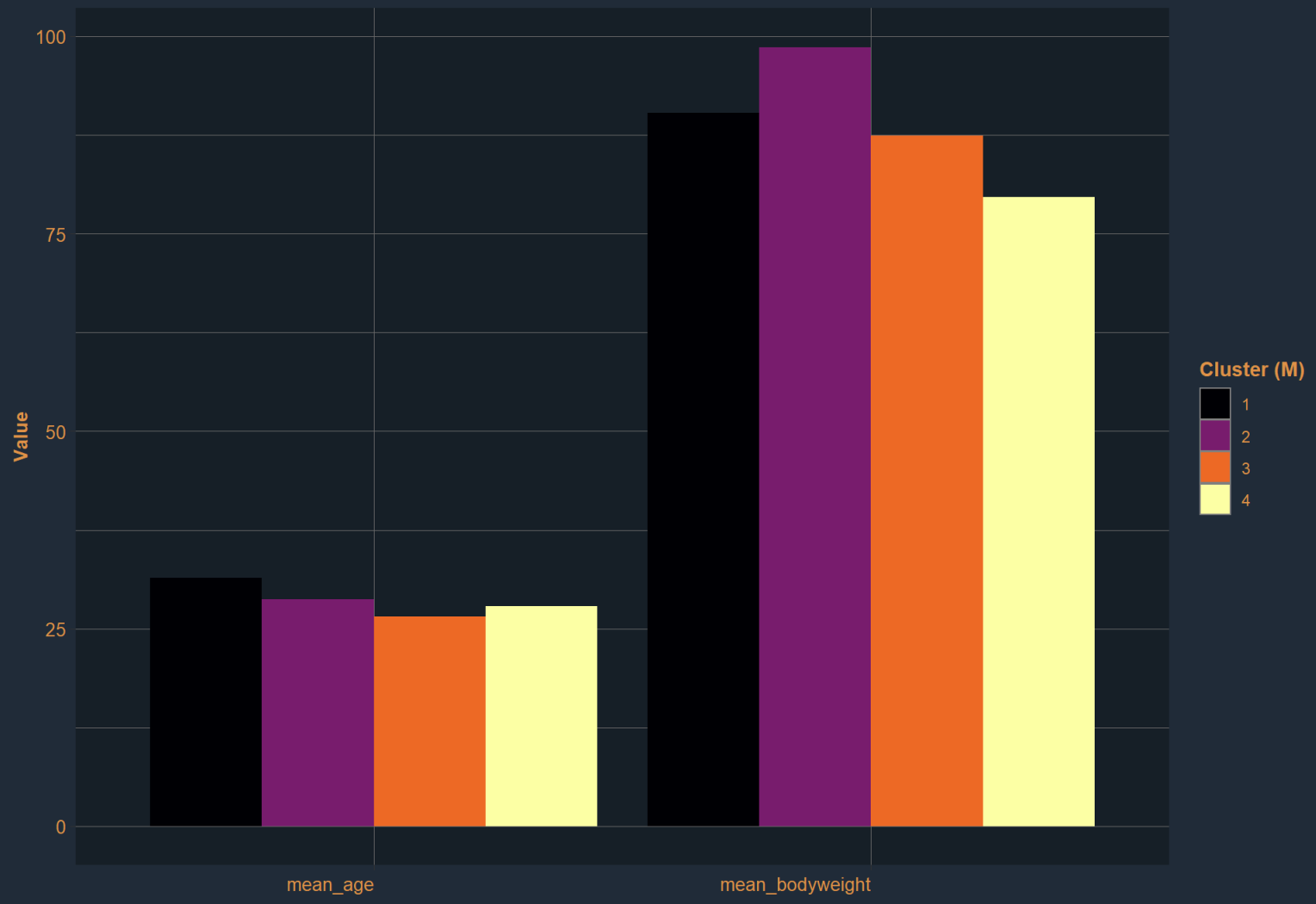
The immediately most noticeable aspect of the women's clusters is probably that they appear more evenly spaced than those of the men. We have neither obtained the clear distinction in the squat, nor that of the bench percentages we could observe in the male population. Instead, both squat and bench percentages each show one group with lower percentage (31% in the squat and 18 in the bench press), two middle-ground groups (34-35% in the squat and around 20 in the bench) and a group of high-performers (38 and 23%). The deadlift percentages are relatively evenly spaced.

Another aspect of note is that the women seem to generally get a lower percentage of their total out of their bench press with a mean of 20.6% across all clusters in comparison to 24.1% across all male groups, while compensating mostly with a noticeably higher deadlift percentage (mean of 44.3% compared to the men's 41.2). The final peculiarity we observed in the men's clusters, the lack of a cluster that obtains high percentages in both bench press and deadlift, also seems to be missing in the women's data, as the group with the second-highest percentage in the bench press, group 3, gets by far the most out of their deadlift with an astounding 48%. In fact, we can notice a different pattern in the women's clusters: The group that obtain a really high percentage of their total in the deadlift, group 3, obtains the smallest amount of their total out of their squat, while group 4 turns this around and gets the most of all groups out of the squat while simulataneously holding the title for smallest deadlift percentage. The relationship of squat to deadlift percentages seems to be to the women what bench to deadlift relationship was for the men's clusters.

Again, I have to mourn the lack of anthropometric data here. It would have been interesting to explore the relationship of height or limb-length with these groups, metrics our dataset unfortunately lacks. What it does not lack, however, are the athlete's totals, wilks, age and bodyweight. So, as the next step, let us take a look at how these metrics vary across clusters!

To start, I will again begin with the male data, illustrating how the athlete's strength, age and bodyweight varied across the different clusters we created. Let's go.

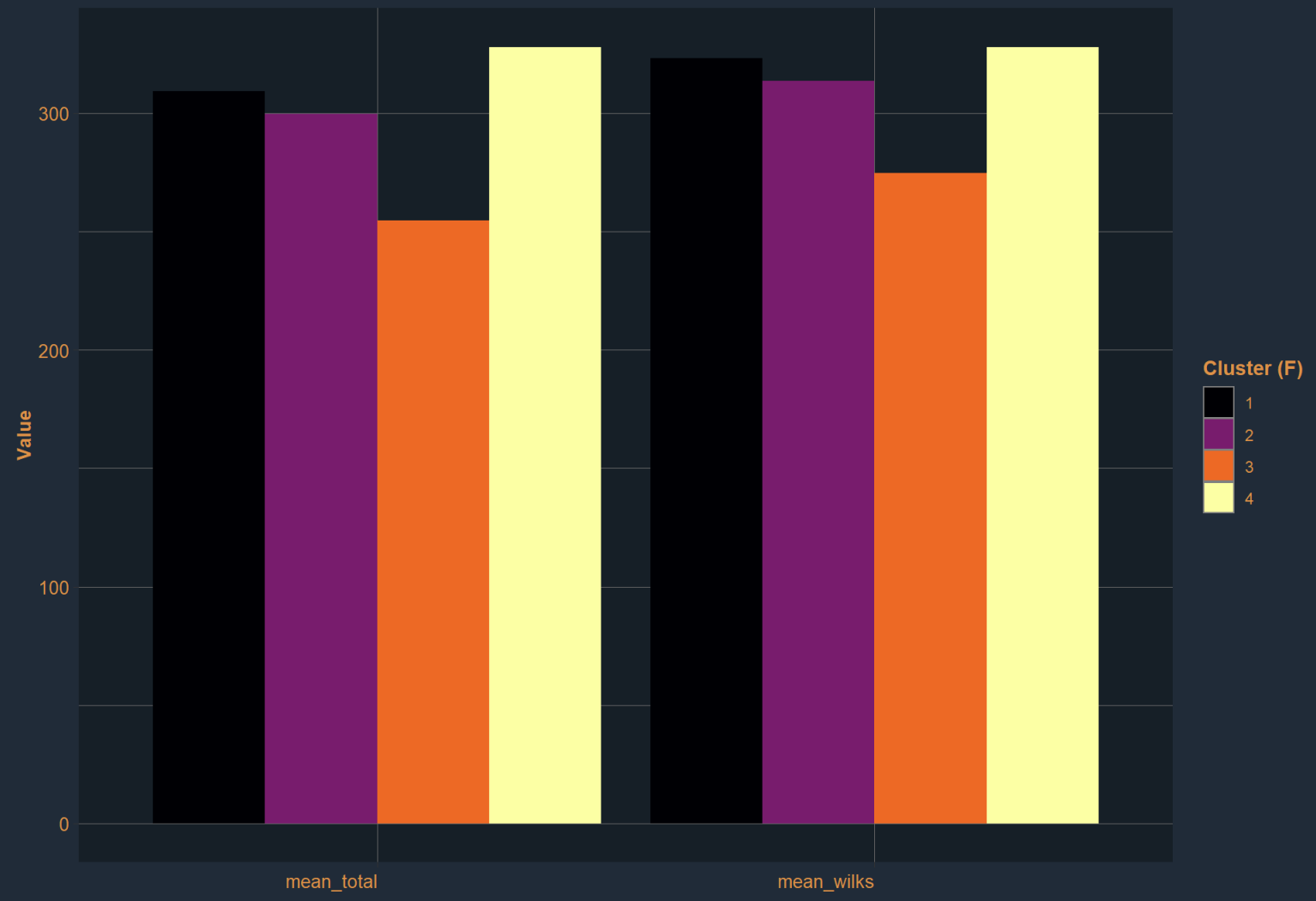


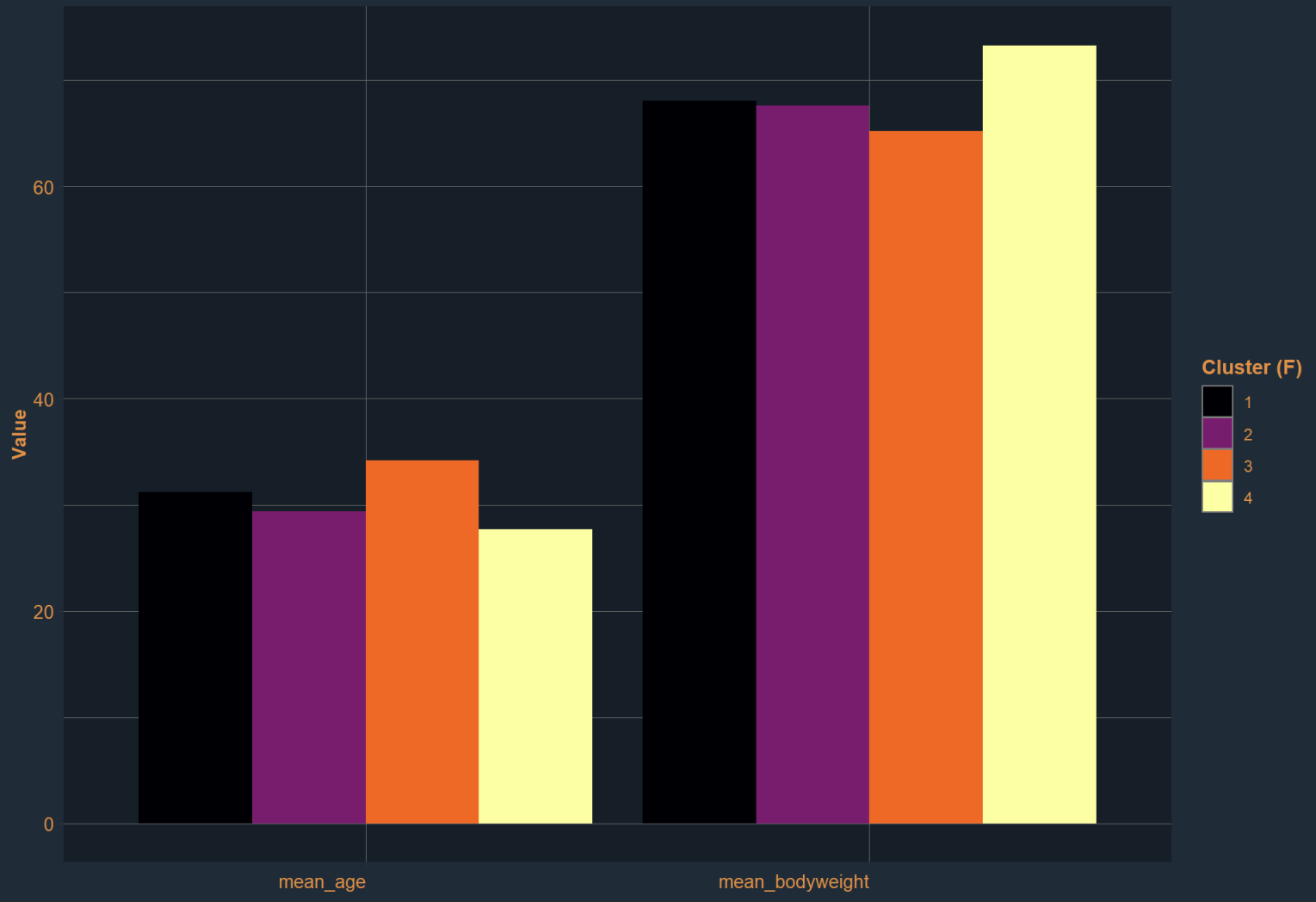


What stands out here is that cluster two is the strongest in the field, sporting the highest mean wilks and total in combination with the highest bodyweight, approaching a mean bodyweight of 100kg. Characteristics of this cluster were high squat and bench percentage, low deadlift percentage (37/25/37).

In contrast to that, cluster four has the lowest wilks and mean total as well as the lowest bodyweight. Characteristics of this group were lower percentages in squat and bench in combination with a very high deadlift percentage (32/22/45). One possible interpretation could be the following: The strongest lifters seem to be able to get roughly the same weight out of squat and deadlift. Maybe this equilibrium can be taken as an indication of a very competent and, as a result, very strong lifter. Excessive deadlift percentage, on the other hand, might be an indication of a more novice lifter. Groups one and three seem to be more of a middle ground, not standing out in any metric.

Next up, let us do the same for the women’s cluters.


300
200
100
0
Value
mean_total
mean_wilks
Cluster (F)
1
2
3
4



What stands out: Cluster three shows up with the lowest total and wilks as well as highest average age. As we saw with the men, this group also sports the highest deadlift percentage while being very low on both squat and bench (31/20/48). This seems to confirm the theory I postulated above: A very high contribution of the deadlift to the total seems to indicate a lifter that still has potential for growth. On the other side of the spectrum, cluster four sports the highest total and wilks as well as the highest bodyweight. Characteristics of this cluster were, again, a relative equilibrium between squat and deadlift percentages (38/20/40). This trend seems to prevail across genders: The lifters in the strongest cluster get roughly the same percentages out of squat and deadlift!

Prediction Tool - Random Forest (Spoiler: Pretty Mediocre)

Now for the final part of this report, I want to try something fun: I want to build a model that tries to use all the various factors and metrics we have explored so far in order to predict an athlete's total.

To get this done, I will use a *random forest*, which can be thought of as a collection of decision trees that tries to split the data at, ideally, the most meaningful 'forks in the road'.

Think about how you would try to make a close estimation of a lifter's total without any prior knowledge and in as few questions as possible: You could start by asking whether the athlete is male or female. Whether he or she competes raw or equipped. Then whether or not they athlete weighs more or less than a certain amount, etc. After a few questions, you would likely have some sort of rough estimate as to what range the total could fall into. If you'd then try this procedure on a number of lifters, you might be able to refine it, steadily finding out which questions actually yield the most information, where exactly certain thresholds are located. You would have developed a *decision tree*, on the basis of which you could come to a reasonable guess about the lifter's total.

However, depending on the sample of athletes you developed your decision tree on, your results may vary. This is why using a collection of decision trees might make sense if we want the results to work best on new lifters, i.e. previously unseen data. This is basically what a *random forest* does. It combines an ensemble of decision trees into an average, which should reduce overfitting and allow for reasonable predictions on new data.

We will make use of the 'BodyweightKG', 'Age', 'Sex', 'Tested', 'Homecountry', 'sq_percentage', 'b_percentage' and 'dl_percentage' columns. It is worth noting that this list does not include the assignment of each athlete to the clusters we dealt with above, even though we saw that they can contain meaningful information about the athletes performance. I chose to instead include the raw material we build the clusters upon in squat, bench and deadlift percentages. I want the model to ideally infer the information these columns contain on itself, instead of relying on the clusters we created.

Now it is time to build the model. We can specify the number of trees we want to go with. We will start with 500 and refine this number later on. We will also set the importance parameter of the model equal to TRUE, since checking which variables the model deems important might prove insightful.

```
random_forest <- randomForest(TotalKg ~ ., data = forest_data,
                              ntree=500,
                              importance = TRUE)
```

```
##
## Call:
##  randomForest(formula = TotalKg ~ ., data = forest_data, ntree = 500,      importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 5157.811
##                     % Var explained: 79.34
```

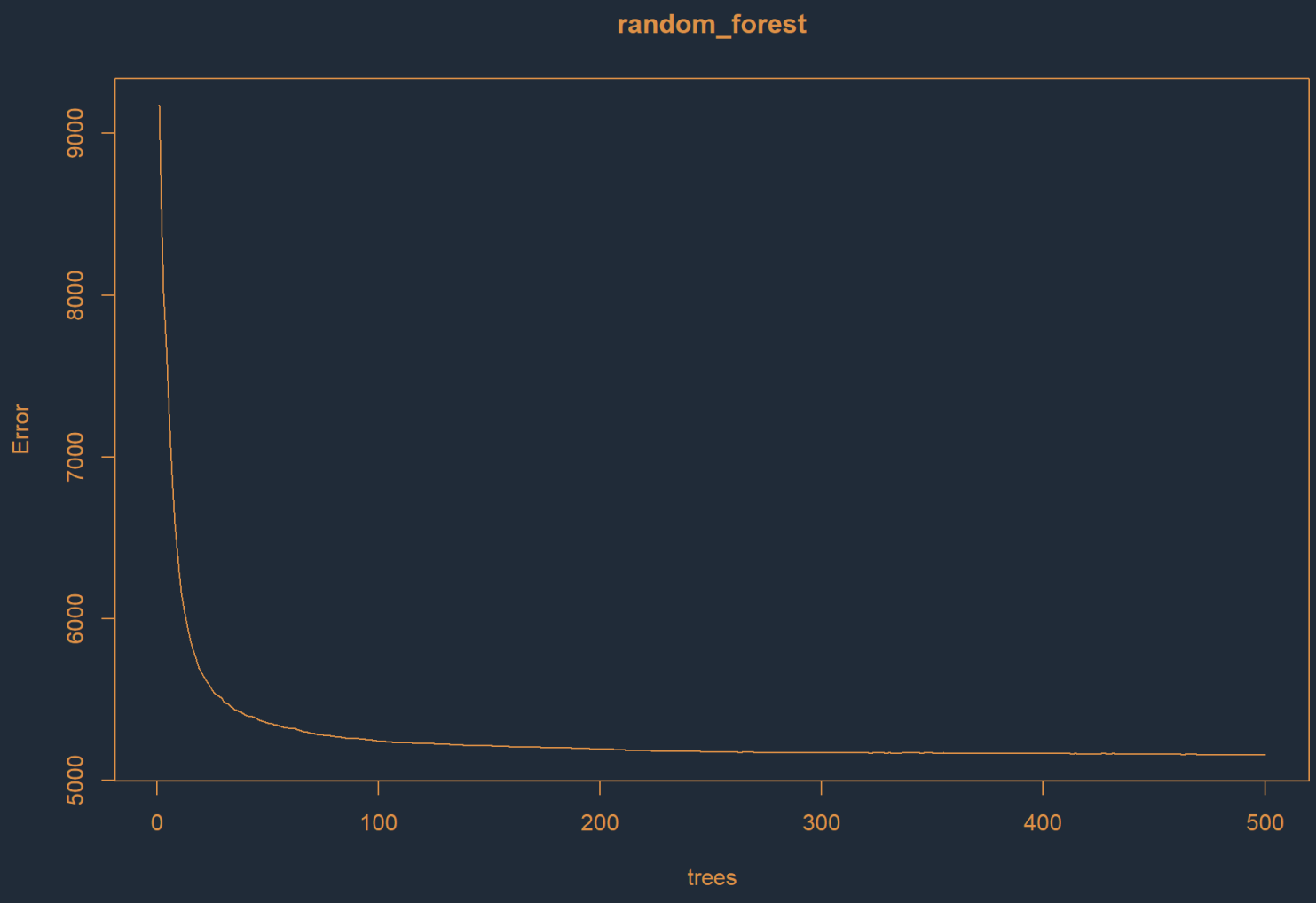
When printing out the models summary, we can see that it was able to explain a little above 79 percent of all variance in the athletes total using the variables we gave it access to. Not too shabby. Let us take a look at what variables the model deemed most important:

```
##                  %IncMSE IncNodePurity
## Tested          120.6232       5301772
```

```
## Sex             14583.8516     360376265
## Age              3081.1782     112307441
## BodyweightKg     8048.2365     338383320
## Homecountry       147.9266       5845657
## sq_percentage    1910.8209      62545739
## b_percentage     1943.7338      95962612
## dl_percentage    2699.0665     130165145
```

These results speak volumes. As we can see in the '%IncMSE' column of the output the percentages of the three lifts, Age, Bodyweight and especially Sex make meaningful contributions to explaining the total. Both the 'Tested' and 'Homecountry' columns seem remarkably unimportant in comparison. Interestingly enough, the three lift's percentages also contribute meaningfully to the reducing error.

Still, we can do better than that! As a first step in tuning the model, let us plot the rate of error in relation to the number of trees we use. This should be helpful in identifying how many decision trees make sense and avoid creating too complex of a model without any real gain in prediction accuracy.



This plot paints a pretty clear picture: The model is able to initially reduce the error quite significantly, but using more than 200 trees seems to hardly have any effect on the error rate at all. so, this will be the first parameter we will adapt in our tuned model: A reduction to only 200 individual decision trees.

Next, let us take a look at the ideal parameter value for 'mtry', the number of variables we randomly sample as candidates at each split in our individual trees. For regression problems, like this one, a good starting point is to use the number of features divided by three, rounded down. Since we are using eight features, that would be an mtry value of 2. To confirm this suspicion, let us use the tunrRF function to take a look at what exactly the optimal parameter value should be based on the observed reduction in error:

```
## mtry = 2  OOB error = 5188.146
## Searching left ...
## mtry = 4     OOB error = 5203.232
## -0.002907888 0.05
## Searching right ...
## mtry = 1     OOB error = 7385.859
## -0.4236029 0.05
```

As we can see, all mtry values between 2 and 4 seem to net us around the same error rate. In addition to all evidence so far pointing in the direction of 2 as the best parameter value, it should be mentioned that a lower value also leads to less correlation between the trees in our random_forest, resulting in a more diverse ensemble. So, let us go with 2, implement both new parameters and see if the performance of our model improves!

```
random_forest_tuned <- randomForest(TotalKg ~ ., data = forest_data,
                                    ntree=200,
```

```
                                mtry=2,
                                importance = TRUE)
```

```
##
## Call:
##  randomForest(formula = TotalKg ~ ., data = forest_data, ntree = 200,      mtry = 2, importanc
##                Type of random forest: regression
##                      Number of trees: 200
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 5189.448
##                     % Var explained: 79.22
```

Well, while the newly tuned model did take significantly less time to fit the data it did not clock in with a noticeable improvement, instead achieving similar results to our model's first version. While an improvement in explained variance was what I was hoping for, it is still nice to see that almost identical results can be achieved using a fraction of the time and computational resources. I'll take it! Let us take a look at the importance of the variables:

```
##                 %IncMSE IncNodePurity
## Tested          122.1802       5469873
## Sex           14664.8819     361996346
## Age            3083.1841     111474782
## BodyweightKg   7991.7458     343781478
## Homecountry     144.0905       5698502
## sq_percentage  1881.6541      65275201
## b_percentage   1915.5386     100498558
## dl_percentage  2622.9048     119412321
```

The contributions of the individual variables to reduction in error look nearly identical to those we previously observed, with only some very slight variations. Seems like the initial model was not too far off in terms of weighting the available variables in terms of their meaning for the prediction of the athlete's totals.

Let us now finally take a look at the predictions the model gives. First off, the models predictions for the first five totals of the data the model was trained on:

```
##        1        2        3        4        5
## 638.5860 280.9457 632.7471 301.2731 520.8012
```

...as well as the actual totals in the data:

```
## [1] 713.50 302.00 721.21 307.50 525.00
```

These predictions do not look too bad! While none of these values hits the nail exactly on the head, all seem to be in a reasonable range around the actual values! Let us take a look at the predictions this model makes for the held-out test data:

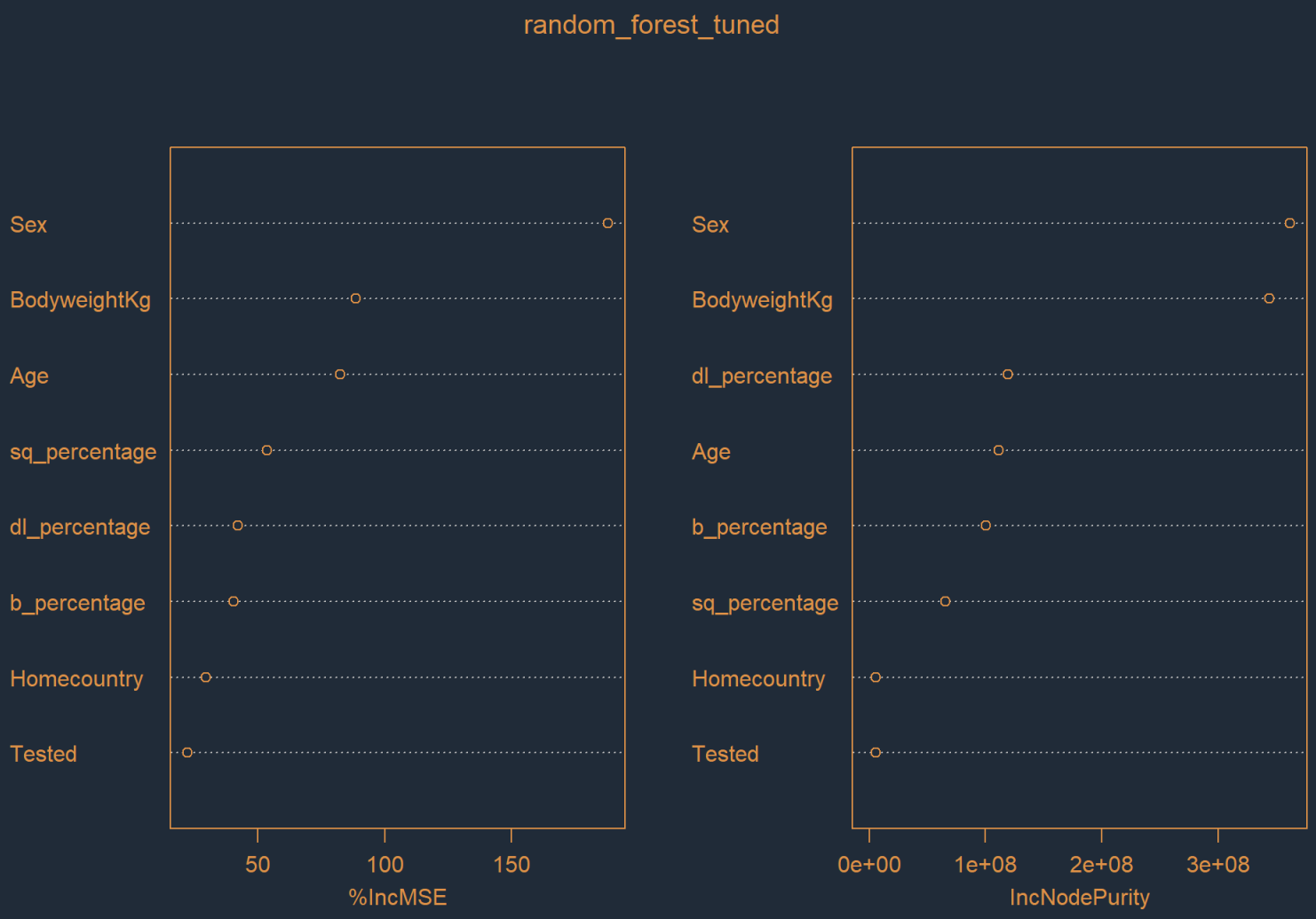
```
##        1        2        3        4        5
## 462.8117 508.1213 294.9690 570.7477 575.9983
```

...and the actual totals:

```
## [1] 505.0 462.5 280.0 667.5 450.0
```

Again, most of the predictions look to he in a decent range around the actual values. Some, especially in the higher ranges of total weight moved, seem to be far less accurate, however.

We have already done this above, but since a visualization might be more intuitive than a chart, we can also visualize the importance of the single variables for the model:



The left table indicates the increase in error we would end up with in case the indicated variable was removed from our pool of predictors. As we can see, Sex, Bodyweight and Age are the most useful in predicting the athlete's total. Still, the three lifts percentages, especially that of the squat, manage to contribute some predictive power as well. Whether or not an athlete competes in their Homecountry and whether or not they competed Tested, however, seem far less relevant, likely indicating that the results we obtained above are partially caused by differences in third factors we did not control for.

The table on the right indicates how much purer, i.e. more aligned the final predictions would end up if the variance from the indicated variable would be taken out of the equation. As already the case in the table on increase of error, we can notice Sex and Bodyweight as the

variables contributing most to node impurity, meaning they serve as the most reliable indicators of where an athlete’s total is at. h

Summary

And this is it. We took a look at how the sport of powerlifting changed during the last two decades, uncovering that there was a massive influx of lifters, during which the percentage of competing women increased substantially.

Meanwhile Raw became the most dominant class of competition, edging out the previously more popular equipped lifting categories.

We took a look at the importance of both age and bodyweight on an athletes strength and modeled the relationship of both factors.

In the second part of the analysis, we found out how significant the differences between tested and untested athletes as well as those competing at home or abroad were (eben though the results should be taken with a grain of salt). We then quantified the difference the different classes of equipment made.

Afterwards, we created clusters of lifters for both the male and female population, comparing them in terms of a few key metrics.

Finally, we built a simple random-forest model for predicting an athlete’s total based on the different variables we explored before. Luckily, the predictions varied in their accuracy, which might very well be taken as an indication that, regardless of age, sex or bodyweight, much of where a lifter ends up lies in his or her own hands.